INTRODUCTION

The estimation of room reverberation time (RT) has been of interest to engineers and acousticians for nearly a century (Sabine, 1922; Kuttruff, 1991). The RT of a room specifies the duration for which a sound persists after it has been switched off. The persistence of sound is due to the multiple reflections of sound from the various surfaces within the room. Historically, the RT has been referred to as the T_{60} time, which is the time taken for the sound to decay to 60 dB below its value at cessation.

Reverberation results in temporal and spectral smearing of the sound pattern, thus distorting both the envelope and fine structure of the received sound. Consequently, the RT of a room provides a measure of the listening quality of the room. This is of particular importance in speech perception where it has been noted that speech intelligibility reduces as the RT increases, due to masking within and across phonemes (Knudsen, 1929; Bolt and MacDonald, 1949; Nabalek and Pickett, 1974; Nabalek and Robinson, 1982; Nabalek *et al.*, 1989). The effect of reverberation is most noticeable when speech recorded by microphones is played back via headphones. Previously unnoticed distortions in the sound pattern are now clearly discerned even by normal listeners (see Hartmann, 1997, for a discussion), highlighting the remarkable echo suppression and dereverberation capabilities of the normal auditory system when the ears receive sounds directly. For hearing impaired listeners, the reception of reverberant signals via the microphone of a hearing aid serves to exacerbate the problem of listening in challenging environments. While derverberation is an active area of investigation, state-of-the-art hearing aids, or other audio processing instruments,

implement signal processing strategies tailored to specific listening environments. These instruments are expected to have the ability to evaluate the characteristics of the environment, and turn on the most appropriate signal processing strategy. Thus, a method that can characterize the RT of a room from passively received microphone signals represents an important enabling technology.

In the early 20th century, Sabine (1922) provided an empirical formula for the explicit determination of RT based solely on the geometry of the environment (volume and surface area) and the absorptive characteristics of its surfaces. Since then, Sabine's reverberation-time equation has been extensively modified and its accuracy improved (see Kuttruff, 1991, for a historical review of the modifications), so that today, it finds use in a number of commercial software packages for the acoustic design of interiors. Formulae for calculation of RT are used primarily for the design of concert halls, classrooms, and other acoustic spaces where the quality of the received sound is of greatest importance, and the extent of reverberation must be controlled. However, to determine the RT of existing environments, both the geometry and the absorptive characteristics have to be first determined. When these cannot be determined easily, it is necessary to search for other methods, such as those based purely on controlled recordings of sounds in the environment to be tested.

Schroeder (1965; 1966) presented a method for estimating RT based solely on the recording of an acoustic signal radiated into the test enclosure. The method obviates some of the problems occurring in situations where the geometry and surface absorption characteristics are unknown. A burst of broad- or narrow-band noise is radiated into the test enclosure until it reaches steady state, and it is then abruptly switched off. The method tracks the sound energy decay following

sound cessation (Schroeder, 1965). This method, referred to as Schroeder's backward integration method, while theoretically and practically important, has some limitations. Specifically, the sound used for measuring RT must be stationary and uncorrelated, and the precise time of sound offset must be known.

Although Schroeder's method has been improved over the years (see Chu, 1978; Xiang, 1995, for example), the improvements do not remove the restrictions placed on the applicability of Schroeder's method. At present, there is no "blind" method that can estimate room RT without requiring knowledge of the geometry, absorptive characteristics, and sound source characteristics or offset time of the sound. A blind method that works with speech sounds would be particularly important for incorporating in hearing-aids or hands-free telephony devices. Partially blind methods have been developed in which the characteristics of the room are "learned" using neural network approaches (Tahara and Miyajima, 1998; Nannariello and Fricke, 1999; Cox *et al.*, 2001), or some form of segmentation procedure is used for detecting gaps in sounds to allow the sound decay curve to be tracked (Lebart *et al.*, 2001). The only other method that may be described as truly blind is "blind dereverberation", in which sound source recovery is attempted by deconvolving the room output with the unknown room impulse response. In principle, this method can be used for extracting the RT, but there are serious drawbacks that limit its applicability. Namely, the room impulse response must be minimum phase, a condition that is rarely satisfied (Neely and Allen, 1979; Miyoshi and Kaneda, 1988).

Here we attempt to address several of the drawbacks found in existing methods by providing

a blind approach that requires only one recording microphone. It does not require that a test signal be radiated into the test enclosure (as in Schroeder's (1965) method) or that the geometry and absorption characteristics of the test enclosure be known (as with the Sabine type formulae). The system performs blind estimation based on a decay curve model describing the reverberation characteristics of the enclosure. Sounds in the test enclosure (speech, music, or other pre-existing sounds) are continuously processed and a running estimate of the reverberation time is produced by the system using a maximum-likelihood parameter estimation procedure. A decision-making step then collects estimates of RT over a period of time and arrives at the most likely RT using an order-statistics filter.

BRIEF DESCRIPTION OF THE FIGURES

FIGURE 1

Temporal decay of a hand–clap at $t = 0.1$ s as recorded by a microphone (left column) and the model matching the reverberation (right column). (A) The recorded sound shows strong early reflections followed by a reverberant tail. Direct sound is excluded from the trace. (B) Model matching the reverberant tail shown in (A). Direct and early reflections are excluded. The model is a Gaussian white noise process damped by a decaying exponential, parametrized by the noise power σ and decay time-constant τ. (C) Decay time-constant estimated from Schroeder's backward integration method (Schroeder, 1965) between -5 dB ($\Diamond$) and -25 dB ($\circ$). Slope of linear fit (dashed line) yields $\tau = 59$ ms ($T_{60} = 0.4$ s). (D) Decay curve for model has identical slope everywhere following sound offset, and captures the most significant part of decay (-5 dB to -25 dB).

FIGURE 2

maximum-likelihood estimation (MLE) of room decay time-constant. (A) The time-constant of the exponential decay (τ, abscissa) is mapped to a parameter $a = \exp(-1/\tau)$ (ordinate) where τ is given in sampling periods. The function is monotone but highly compressive and maps $\tau \in [0, \infty)$ onto $a \in [0, 1)$. Filled circle shows $\tau = 100$ ms ($a = 0.9994$). (B) Score function (derivative of log likelihood function) $s_a(a)$ (ordinate), decreases rapidly as a function of a (abscissa, marked in time constants using the map in (A)). MLE of a is given by the root of $s(a)$ (filled circle). (C) The derivative $s_a\prime(a)$ as a function of a. At the root of s_a (filled circle), the derivative is negative. Note the nearly 8–12 orders of magnitude change in s_a and $s_a\prime$ for commonly encountered values of τ. (D) The ratio $s_a(a)/s_a\prime(a)$ (ordinate) as a function of a is the incremental step size of the Newton–Raphson procedure for finding the root of Eq. (8). It provides an estimate of the convergence properties of the root-finding algorithm. Sampling frequency was 16 kHz, and the log-likelihood function was calculated assuming a 400 ms window.

FIGURE . 3

Illustration of procedure for continuous estimation of room decay time. A burst of white noise was applied at time $t = 0.1$ s (black bar, bottom trace, 100 ms duration). Simulated room output (bottom trace) shows the build–up and decay of sound in the room. A running estimate of the parameter a in 200 ms windows is shown in the graph (ordinate, a shown in units of time-constant). The true value of room decay time (100 ms) is shown as horizontal dashed line. The estimation window was advanced by one sample to obtain the trace, with each point at time t being the estimate in the window $(t - 0.2, t]$. During the build-up and ongoing phase of the sound, estimated a sometimes exceeded 1 (i.e., negative values of τ). These were discarded and are not shown. As the window moved into the region of sound decay ($t > 0.3$ s), the estimates converged to the correct value. A histogram of the estimated time-constant is shown to the right of the trace. An order–statistics filter, such as the mode of the histogram, can be used to extract the room time-constant. Sampling rate was 16 kHz.

FIGURE 4

Effect of estimation window length on the variance of the estimate. The simulation shown in Fig. 3 was repeated for windows of duration 0.5τ, τ, 2τ, and 4τ (top to bottom), where $\tau = 100$ ms is the true value of the room time-constant. The left column shows the running estimate of parameter a (ordinate, shown as time-constant in ms) as a function of time (abscissa). The right column shows the histogram of the estimates. The variance of the estimate decreases with increasing window length (arrowheads mark true value of τ).

FIGURE 5

Estimation of room time-constant from speech. Fifteen words recorded in an anechoic (clean) environment (200 ms inter–word spacing) were convolved with a room model (τ = 100 ms). Histograms of decay time-constants were estimated from clean (left column) and simulated reverberant responses (right column), and are shown for window durations 0.5τ, τ, 2τ, and 4τ (top to bottom). The histogram for clean speech served as a control. Description follows Fig. 4. Estimation from reverberant speech produces a clearly defined peak, especially for the longer window lengths, albeit with a small bias (right column, 2τ and 4τ). The bias can be attributed to the gradually decaying offsets inherent in speech so that the resultant decay is a convolution of speech offset and the room response. For the control condition (left column), the offset decay is visible only in the bottom two rows where the histogram exhibits a broad bump between 50 and 100 ms. The fifteen words included 11 /p,b,g/V/d/ and 4 /b/V/ sampled at 20 kHz.

FIGURE 6

Illustration of decay-time estimation when a terminating phoneme is encountered. The word "bough" recorded under anechoic (clean) conditions (top row) has a gradually decaying offset. The envelope was extracted by filtering the absolute value of the analytic signal (second row, heavy outline), and its decay rate was estimated for the segment following the dashed line using two methods (bottom row). Overlapping segments (duration given by bar, with step size indicated by the thickness of the vertical end) were converted to a decibel scale and the decay time-constant obtained by a least squares fit to a straight line (dotted trace). The same segments were analyzed using the MLE algorithm to obtain a second estimate of the decay time-constant (solid trace). While the estimators provide somewhat different results, they are in qualitative agreement. Both methods suggest that the fastest decay time-constant is in the range of 50-70 ms (see also Fig. 5). These results suggest that speech segments will introduce a bias when estimation is carried out in reverberant environments.

FIGURE 7

Comparison of RT estimates obtained from MLE method and Schroeder's method. (A) Mean RT (ordinate) in one-third octave bands (abscissa) averaged over 100 independent trials of a simulated decay curve (RT = 0.5 s). RT estimates were obtained using the MLE procedure (circles), and Schroeder's method in 20 dB decay range (lozenge), and 30 dB decay range (square). The filled symbols are broad-band estimates. (B) Standard deviation of the RT for broad-band and one-third octave bands over 100 trials. Symbols follow (A).

FIGURE 8

Estimation of decay time-constant from real room data. (A) The room response to a hand-clap (same as Fig. 1A but includes the direct sound). (B) Spectrogram of the hand-clap demonstrates a sharp broadband onset transient and the decay as a function of frequency. (C) The decay time-constant was estimated using Schroeder's backward impulse integration method in the -5 dB (lozenge) to -25 dB (circle) range, followed by a least-squares fit to a straight line to obtain the time-constant (τ = 56 ms, T_{60} = 0.39 s). (D) Histogram of decay times obtained from signal shown in A using MLE. The median value of the histogram (arrow, τ = 53 ms, T_{60} = 0.37) is in good agreement with the estimate obtained using Schroeder's method.

FIGURE 9

Comparison of Schroeder's method and the MLE procedure for T_{60} times obtained in one-third octave bands. Three environments were tested: a moderately reverberant environment (circles; the environment is the same as shown in Fig. 8), a highly reverberant circular foyer (squares), and a highly reverberant enclosed cafeteria (diamonds). In each environment, a single hand-clap was filtered using a bank of ISO one-third octave band-pass filters with center frequencies exceeding 1 kHz. The ordinate shows the best estimates obtained from the MLE procedure for each band, and the abscissa shows the T_{60} times obtained from Schroeder's method. Averages over all bands for each environment are shown as filled symbols. The diagonal dashed line (with unity slope) is shown for reference, and points lying close to this line suggest good agreement between the two procedures. Agreement is best when the T_{60} values are averaged over all the bands.

FIGURE 10

Reverberation-time estimates from real environments. Seventeen tests in 12 environments were conducted using noise bursts. Decay time-constants were estimated using the MLE algorithm (ordinate) and the extrapolated T_{60} times were compared with estimates from Schroeder's method (abscissa). (A) Estimates of T_{60} in one-third octave bands with center frequencies exceeding 1 kHz (open circle) and their average (filled square). (B) Broad-band estimates of T_{60} from the recorded room response. Averaged narrow-band estimates are more accurate than broad band estimates, presumably due to the presence of low-frequency components in the latter. Further, the MLE method over-estimates T_{60} (in comparison with Schroeder's method) when room reverberation is moderate (< 0.3 s), whereas for higher values of T_{60} (> 1.3 s) there is reasonable agreement. The single outlier is due to inaccuracies in the Schroeder estimate (see text for discussion). Results are from one presentation of noise burst in each test.

FIGURE 11

Evaluation of room reverberation time (RT) from connected speech played back in a partially open circular foyer. The RT for this environment as measured from hand-claps was 1.66 ± 0.07 s (Schroeder's method) and 1.62 s (from ML procedure). (A) Trace of CST passage (duration 50 s) recorded in the environment. Bar indicates 1 s. (B) The histogram of ML estimates over the duration of recording. The first peak in the aggregate histogram is the best RT estimate from connected speech (1.83 s). The horizontal bar is the range of RT estimates obtained from Schroeder's method, and the triangle indicates the ML estimate. (C) Peak values from histogram of estimates were obtained every 1 s, and the 50 peak values were used to produce the histogram shown. The best estimate of RT from this histogram is at the dominant peak (1.7 s), which is closer to the estimates obtained from pure decay curves. Thus, using short-term histograms as in (C) is more reliable than the long-term histogram shown in B. Overall, the results indicate that the ML estimator produces reliable estimates with connected speech.

FIGURES 12-23

Figures 12-23 provide charts and graphs associated with various aspects of reverberation evaluation described herein.

FIGURE 24

Figure 24 is a diagrammatic view of one system for implementing reverberation evaluation.

DETAILED DESCRIPTION

While the present invention can take many different forms, for the purpose of promoting an understanding of the principles of the invention, reference will now be made to the embodiments illustrated in the drawings and specific language will be used to describe the same. It will nevertheless be understood that no limitation of the scope of the invention is thereby intended. Any alterations and further modifications of the described embodiments, and any further applications of the principles of the invention as described herein are contemplated as would normally occur to one skilled in the art to which the invention relates.

Figure 24 illustrates system 20 of one embodiment of the present invention. System 20 is configured to detect sound with sensor 22 emanating from one or more acoustic sources 24 in room 26. Sensor 22 generates a corresponding sensor signal representative of the detected sound. For the example illustrated, only one sensor 22 is shown; however, more than one sensor may be utilized. Sensor 22 can be in the form of an omnidirectional dynamic microphone or a different type of microphone or sensor type as would occur to one skilled in the art. Sources 24 can be actively controlled and/or passive in nature. Indeed, in accordance with the teachings of the present invention, reverberation of an acoustic environment can be evaluated based only on the processing of sensor signals representative of uncontrolled, passive sound emanating from such environment.

System 20 further includes processing subsystem 30. Subsystem 30 includes at least one processor 32 and memory 34. Memory 34 includes removable memory device 36. Sensor 22 is operatively coupled to processing subsystem 30 to process signals received therefrom. Processing subsystem 30 is operable to provide an output signal

representative of acoustic excitation detected with sensor 22 that may be modified in accordance with processing routines and/or parameters of subsystem 30. This output signal is provided to one or more output devices 40. In figure 24, the one or more output devices 40 are labeled in the plural, but is also intended to be representative of the presence of only a single output device. In one embodiment, at least of one output devices 40 presents an output to a user in the form of an audible or visual signal. In other embodiments, at least one of output devices 40 provides a different user/operator output and/or is in the form of other equipment that utilizes the output signal for further processing. In still other embodiments, one or more output devices 40 are absent (not shown).

Processor 32 is responsive to signals received from sensor 22. Processor 32 can be of an analog type, digital type, or a combination of these. Subsystem 30 can include appropriate signal conditioning/conversion to provide an appropriate sound-representative signal to processor 32 from sensor 22. Processor 32 may be a software or firmware programmable device, a state logic machine, or a combination of both programmable and dedicated hardware. Furthermore, processor 32 can be comprised of one or more components and/or can include one or more independently operable processing components. For a form with multiple independently operable processing components; distributed, pipelined, and/or parallel processing can be utilized as appropriate. In one embodiment, processor 32 is in the form of a digitally programmable signal processing semiconductor component that is highly integrated. In other embodiments, processor 32 may be of a general purpose type or other arrangement as would occur to those skilled in the art.

Likewise, memory 34 can be variously configured as would occur to those skilled in the art. Memory 34 can include one or more types of solid-state electronic memory, magnetic memory, or optical memory of the volatile and/or nonvolatile variety. Furthermore, memory can be integral with one or more other components of processing subsystem 30 and/or comprised of one or more distinct components. Memory 34 can be at least partially integrated with processor 32. Removable memory device 36 is of a computer/processor accessible type that is portable, such that it can be used to transport data and/or operating instructions to/from subsystem 30. Device 36 can be of a floppy disk, cartridge, or tape form of removable electromagnetic recording media; an optical disk, such as a CD or DVD type; an electrically reprogrammable solid-state type of nonvolatile memory, and/or such different variety as would occur to those skilled in the art. In one embodiment, device 36 is utilized to load and/or store at least a portion of the operating logic for subsystem 30. This operating logic can be in the form of instructions carried by device 36 that are executed by processor 32 to perform one or more routines according to the present invention. In other embodiments, some or all of this operating logic is stored in another portion of memory 34 and/or is defined by dedicated logic of subsystem 30 and/or processor 32. In still other embodiments, device 36 is absent.

Processing subsystem 30 can include one or more signal conditioners/filters 32a to filter and condition input signals and/or output signals; one or more format converters, such as Analog-to-Digital (A/D) and/or Digital-to-Analog (DAC) converter types; and/or one or more oscillators, control clocks, interfaces, limiters, power supplies, communication ports, or other types of components/devices as would occur to those

skilled in the art to implement the present invention. In one embodiment, subsystem 30 is provided in the form of a single microelectronic device.

System 20 can be implemented in any of a number of various ways in different embodiments. By way of nonlimiting example, system 20 can be utilized in hearing assistance devices for the hearing impaired and/or for surveillance. In other embodiments, system 20 is utilized in speech recognition arrangements, hands-free telephony devices, remote telepresence or teleconferencing configurations, sound level evaluation equipment, or different applications as would occur to those skilled in the art. In all these embodiments, the evaluation of one or more reverberation characteristics of a subject acoustic environment, such as a room or outside region, is often desired to improve performance.

System 20, and the various embodiments, variations, and forms described in connection with system 20, are but a few examples of arrangements that can be used to implement the reverberation evaluation techniques described in the text accompanying figures 1-23 as follows.

A model for blind estimation of reverberation time is presented. This is followed by an algorithm for implementation, and a decision-making strategy for selecting the estimate that best represents the reverberation time of listening rooms.

Before describing the model, we motivate the work with an example. The recorded response of a room to an impulsive sound source (a hand-clap) is shown in Fig. 1A. As can be expected, there are strong early reflections followed by a decaying reverberant tail. If the early reflections are ignored, the decay rate of the tail can be estimated from the envelope. A common and widely used measure of the decay time is the T_{60} time first defined by Sabine (1922), which measures the time taken for the sound level to drop 60 dB below the level at sound cessation. In practice, a decaying

sound in a real environment reaches the ambient noise floor, thus limiting the dynamic range of the measured sound to values less than 60 dB, and so it is not usually possible to directly measure T_{60}. Instead, the time to reach -25 dB or -35 dB from a reference level of -5dB is often used (Schroeder, 1965). These values can be extrapolated to obtain T_{60}. Figure 1C shows the measurement of T_{60} from the hand-clap data using Schroeder's method (Schroeder, 1965) described below. Schroeder's method suffers from the drawback that the precise instant of cessation of sound must be known, and there must be a sufficiently long period of silence to perform the estimation. Thus, it is not amenable to online implementation when sounds such as connected speech are present.

We begin with a model for the diffusive or reverberant tail of sounds in a room. This refers to the dense reflections that follow the early reflections. All that can be said about them is that they are the result of multiple reflections, and appear in random order, with successive reflections being damped to a greater degree if they occur later in time. Traditionally, and dating back to Sabine, the decay envelope has been modeled as an exponential with a single time-constant. Because the dense reflections are assumed to be uncorrelated, a convenient though highly simplified model is to consider the reverberant tail to be an exponentially damped uncorrelated noise sequence with Gaussian characteristics. The model does not include the direct sound or early reflections. The goal is to estimate the time-constant of the envelope.

Model of sound decay

We assume that the reverberant tail of a decaying sound y is the product of a fine structure x that is random process, and an envelope a that is deterministic. A central assumption is that x is a wide-band process subject to rapid fluctuations, whereas the variations in a are over much longer time-scales. Here, we will provide a statistical description of the reverberant tail with the goal of estimating the decay time-constant of the envelope.

Let the fine structure of the reverberant tail be denoted by a random sequence $x(n)$, $n \geq 0$ of independent and identically random variables drawn from the normal distribution $\mathcal{N}(0, \sigma)$. Further, for each n we define a deterministic constant $a(n) > 0$. The model for room decay then suggests that the observations y are specified by the sequence $y(n) = a(n)x(n)$. Due to the time-varying term $a(n)$, the $y(n)$ are independent but not identically distributed, and their probability density function is $\mathcal{N}(0, \sigma a(n))$. That is, the constant $a(n)$ modulates the instantaneous power of the fine structure. For purposes of estimating the room decay time, we consider a finite sequence of observations, $n = 0, \ldots, N-1$ where N will be referred to as the estimation interval, or estimation window length. For notational simplicity, denote the N-dimensional vectors of y and a by $\boldsymbol{y}$ and $\boldsymbol{a}$, respectively. Then the likelihood function of $\boldsymbol{y}$ (the joint probability density), parameterized by $\boldsymbol{a}$ and σ, is

$$L(\boldsymbol{y};\, \boldsymbol{a},\, \sigma) = \frac{1}{a(0)\ldots a(N-1)} \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{\sum_{n=0}^{N-1}(y(n)/a(n))^2}{2\sigma^2}\right), \qquad (1)$$

where $\boldsymbol{a}$ and σ are the $(N+1)$ unknown parameters to be estimated from the observation $\boldsymbol{y}$. The likelihood function given by Eq. (1) is somewhat general, and while it is possible to develop a procedure for estimating all $(N+1)$ parameters, suitable simplifications can be made when modeling sound decay in a room. Let a single time-constant τ describe the damping of the sound envelope during free decay. Then the sequence $a(n)$ is uniquely determined by

$$a(n) = \exp(-n/\tau). \tag{2}$$

Thus, the N–dimensional parameter $\boldsymbol{a}$ can be replaced by a scalar parameter a that is expressible in terms of τ and a single parameter $a = \exp(-1/\tau)$, so that

$$a(n) = a^n. \tag{3}$$

Introducing Eq. (3) into Eq. (1) yields

$$L(\boldsymbol{y};\, a,\, \sigma) = \left(\frac{1}{2\pi a^{(N-1)}\sigma^2}\right)^{N/2} \exp\left(-\frac{\sum_{n=0}^{N-1} a^{-2n}\, y(n)^2}{2\sigma^2}\right). \tag{4}$$

For a fixed observation window N and a sequence of observations $y(n)$, the likelihood function is parameterized solely by the time-constant a and the diffusive power σ.

The model is shown in Fig. 1B with parameters a and σ matched to the experimental hand-clap data shown in Fig. 1A. Note that the model does not include the early reflections shown in panel A.

The Schroeder decay curve for the model is shown in Fig. 1D with a T_{60} time of 0.4 s in agreement with the measured T_{60}. The agreement between model and real T_{60} time motivates the search for an algorithm that can optimally estimate the two parameters.

Maximum-likelihood estimation

Given the likelihood function, the parameters a and σ can be estimated using a maximum-likelihood approach (Poor, 1994). First, we take the logarithm of Eq. (4) to obtain the log–likelihood function

$$\ln L(\boldsymbol{y};\, a,\, \sigma) = -\frac{N(N-1)}{2}\ln(a) - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1} a^{-2n}\, y(n)^2. \tag{5}$$

To find the maximum of $\ln(L)$, we differentiate the log–likelihood function Eq. (5) with respect to a to obtain the score function (Poor, 1994)

$$\begin{aligned} s_a(a;\, \boldsymbol{y},\, \sigma) &= \frac{\partial \ln L(\boldsymbol{y};\, a,\, \sigma)}{\partial a}, & (6)\\ &= -\frac{N(N-1)}{2a} + \frac{1}{a\sigma^2}\sum_{n=0}^{N-1} n\, a^{-2n}\, y(n)^2. & (7) \end{aligned}$$

The log-likelihood function achieves an extremum when $\partial \ln L(\boldsymbol{y};\, a,\, \sigma)/\partial a = 0$; that is, when

$$-\frac{N(N-1)}{2a} + \frac{1}{a\sigma^2}\sum_{n=0}^{N-1} n\, a^{-2n}\, y(n)^2 = 0. \tag{8}$$

The zero of the score function provides a best estimate in the sense that $\mathbf{E}[s_a] = 0$.

Denote the zero of the score function s_a, and satisfying Eq. (8), by a^*. It can be shown that the second derivative $\frac{\partial^2 \ln L(\boldsymbol{y}; a, \sigma)}{\partial a^2}|_{a=a^*} < 0$, i.e., the estimate a^* maximizes the log–likelihood function.

The diffusive power of the reverberant tail, or variance σ^2, can be estimated in a similar manner. Differentiating the log-likelihood function Eq. (5) with respect to σ, we have

$$s_\sigma(\sigma;\, \boldsymbol{y},\, a) = \frac{\partial \ln L(\boldsymbol{y};\, a,\, \sigma)}{\partial \sigma}, \tag{9}$$

$$= -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{n=0}^{N-1} a^{-2n} y(n)^2, \tag{10}$$

which achieves an extremum when

$$\sigma^2 = \frac{1}{N} \sum_{n=0}^{N-1} a^{-2n} y(n)^2. \tag{11}$$

As before, it can be shown that the $\mathbf{E}[s_\sigma] = 0$. Denote the zero of the score function s_σ, and satisfying Eq. (11), by σ^*. It can be shown that the second derivative $\frac{\partial^2 \ln L(\boldsymbol{y}; a, \sigma)}{\partial \sigma^2}|_{\sigma=\sigma^*} < 0$, i.e., the estimate σ^* maximizes the log–likelihood function. Note that the maximum-likelihood equation given by Eq. (8) is a transcendental equation and cannot be inverted to solve directly for a^*, whereas the solution of Eq. (11) for σ^* is direct. This is considered in detail later when an algorithm for estimation is developed.

Bounds on the estimate of a and σ are obtained from the variance of the score function, also called the Fisher information J. This is more conveniently expressed in terms of the derivatives of the score functions (Poor, 1994). Given the parameter $\theta^{\mathrm{T}} = [a\ \sigma]$ and the score function $s_\theta^{\mathrm{T}}(y;\ \theta) = [s_a(y;\ a,\ \sigma)\ s_\sigma(y;\ a,\ \sigma)]$, we have

$$J(\theta) = -\mathrm{E}[\frac{\partial\, s_\theta^{\mathrm{T}}(y;\ \theta)}{\partial \theta}]. \tag{12}$$

From Eq. (7), Eq. (9), and Eq. (12), we have

$$J(\theta) = \begin{pmatrix} \frac{N(N-1)(2N-1)}{3a^2} & \frac{N(N-1)}{a\sigma} \\ \frac{N(N-1)}{a\sigma} & \frac{2N}{\sigma^2} \end{pmatrix}. \tag{13}$$

By the Cramer–Rao theorem (Poor, 1994), a lower bound on the variance of any unbiased estimator is simply $J^{-1}(\theta)$, which is

$$J^{-1}(\theta) = \begin{pmatrix} \frac{6a^2}{N(N^2-1)} & -\frac{3a\sigma}{N(N+1)} \\ -\frac{3a\sigma}{N(N+1)} & \frac{\sigma^2(2N-1)}{N(N+1)} \end{pmatrix}. \tag{14}$$

From the asymptotic properties of maximum-likelihood estimators (Poor, 1994), we know that the estimates of a and σ are asymptotically unbiased and their variances achieve the Cramer-Rao lower bound (i.e., they are efficient estimates). Thus, if a^* and σ^* are the estimates obtained from the solutions of Eq. (8) and Eq. (11), the variance of the estimates are

$$\mathrm{E}[(a^* - a)^2] \geq \frac{6a^2}{N(N^2-1)}, \tag{15}$$

$$\mathrm{E}[(\sigma^* - \sigma)^2] \geq \frac{\sigma^2(2N-1)}{N(N+1)}, \tag{16}$$

with equality being achieved in the limit of large N. As the variance of a and σ are $\mathrm{O}(N^{-3})$ and $\mathrm{O}(N^{-1})$, the estimation error can be made arbitrarily small if observation windows are made sufficiently large.

Algorithm for estimating decay time

Given an estimation window length and the sequence of observations $y(n)$ in the window, the zero of the score function Eq. (8) provides an estimate of a. The function is a transcendental equation that must be solved numerically using an iterative procedure. However, the estimate of σ can be obtained directly from Eq. (11). A two-step procedure was followed: (1) an approximate solution for a^* from Eq. (8) was obtained, and (2) the value of σ^* was updated from Eq. (11). The procedure was repeated, providing successively better approximations to a^* and σ^*, and so converging on the root of Eq. (8).

Here we address the strategy for extracting the root in the smallest number of iterative steps. To gain an understanding of the root-solving procedure, we consider the example shown in Fig. 2. The function $a = \exp(-1/\tau)$ maps the room time-constant τ one-to-one and onto a as shown in

Fig. 2A. For instance, consider a room time-constant of 0.1 s and a sampling rate of 16 kHz. Then the time-constant is 1600 samples, and so $a = 0.9994$ (filled circle). The signficance of the number becomes clear if we consider that when $\tau = 0.03\ s$, then $a = 0.9979$, whereas for $\tau = \infty$, $a = 1$. Hence the geometric ratio is highly compressive and values of a for real environments are likely to be close to 1. Thus, the advantage of estimating a rather than τ is due to the bounded nature of a. The score function s_a from Eq. (7) on the other hand, has a wide range (about 8 orders of magnitude, see Fig. 2B) and is zero at the room time-constant (filled circle). The gradient of the score function ds_a/da shown in Fig. 2C also demonstrates a wide range, but takes a negative value at the zero of s_a.

Thus, if we start with an initial value of $a_0^* < a$, the root-solving strategy must descend the gradient sufficiently rapidly. The standard method for solving this kind of nonlinear equation, where an explicit form for the gradient is available, is the Newton-Raphson method which offers second-order convergence (Press et al., 1992). The order of convergence can be assessed from $s_a\,(ds_a/da)^{-1}$ which is the incremental step size Δa in the iterative procedure (Fig. 2D). For example, with true value of $\tau = 100$ ms, Δa at intermediate values in the iteration can be as small as 10^{-6} when $a = 0.9993$ ($\tau = 90$ ms) or $a = 0.9995$ ($\tau = 120$ ms). This corresponds to an incremental improvement of about 0.01 ms for every iteration, thus providing slow convergence if the initial value is far from the zero. On the other hand, the bisection method (Press et al., 1992) guarantees rapid gradient descent but works poorly in regions where the gradient changes relatively

slowly (such as near the true value of a). Furthermore, it guarantees only first-order convergence.

However, the specific structure of the root-solving problem can be exploited because the behavior of s_a is known. Here, both methods were used to obtain rapid convergence to the root. First, the root was bisected until the zero was bracketed, after which the Newton-Raphson method was applied to polish the root. For the example shown, the root bracketing was accomplished in about 8 steps and the root polishing in 2-4 steps. In contrast, with the same initial conditions, the Newton-Raphson method took about 500 steps to converge. Taken together, the analysis presented here suggests that the estimation procedure is feasible and does not lead to significant errors although values of a for real rooms are close to 1, and the score function and its derivative vary over many orders of magnitude. While other root-solving procedures are possible, such as iterative gradient optimization, these are not dealt with here.

Strategy for assigning the correct decay time from the estimates

The theory presented in the preceding section provides one estimate of a and σ in a given time frame of N samples. By advancing the frame as the signal evolves in time, a series of estimates a_k^* will be obtained, where k is the time frame. Some of these estimates will be obtained during a free decay following the offset of a sound segment (correct estimations), whereas some will be obtained when the sound is ongoing (incorrect estimations due to model failure). Thus, a strategy is required for selecting only those estimates that correctly represent regions of free-decay and hence the real room time-constant. This requires a decision-making strategy that examines the distribution of the

estimates after a sufficient number of frames have been processed, and makes a decision regarding the true value of the room time-constant.

In a blind estimation procedure the input is unknown, and so the model will fail when: (1) an estimate is obtained in a frame that is not ocurring during a free decay. This includes regions where there is sound onset or sound is ongoing. In these periods, the MLE scheme can provide widely fluctuating or implausible estimates due to model failure. (2) During a region of free decay initiated by a sound with a gradual rather than rapid offset. In this case, the offset decay of the sound will be convolved with the room response, prolonging the sound even further and so, the estimated time-constant will be larger than the real room time-constant. Gradual offsets occur in many natural sounds, such as terminating vowels in speech. We address both issues here and provide a strategy for selecting the correct room time constant.

In the first case where the estimation frames do not fall within a region of free decay, many of the time frames will provide estimates of a close to unity (i.e., infinite τ), or implausible values. On the other hand, the estimates will accurately track the true value when a free decay occurs. Intuitively, a strategy for selecting a from the sequence a_k^* is guided by the following observation: the damping of sound in a room cannot occur at a rate *faster* than the free decay, and thus all estimates a^* must attain the true value of a as a lower bound. The bound is achieved only when a sound terminates abruptly, upon which the model conditions will be satisfied, and the estimator will track the true value of the time-constant.

Although it seems intuitive to set $a = \min\{a_k^*\}$, it should be recognized that even during a

free decay the estimate is inherently variable (due to the underlying stochastic process), and so selecting the minimum is likely to underestimate a.

A robust strategy would be to select a threshold value of a^* such that the left tail of the probability density function of a^*, $p(a^*)$, occupies a pre–specified percentile value γ. This can be implemented using an order statistics filter specified by

$$a = \arg\{P(x) = \gamma : P(x) = \int_0^x p(a^*)\, da^*\}. \tag{17}$$

For a unimodal symmetric distribution with $\gamma = 0.5$ the filter will track the peak (median) value. Order statistics filters play an important role in robust estimation, especially when data is contaminated with outliers (Pitas and Venetsanopoulos, 1992), as is the case here. It should be noted that for γ values approaching 0, the filter Eq. (17) performs like the minimum filter $a = \min\{a_k^*\}$ suggested above.

In the second case described above, where the sound offset is gradual, $p(a^*)$ is likely to be multimodal because sound offsets (such as terminating phonemes in speech) will have varying rates of decay, and their presence will give rise to multiple peaks. The strategy then is to select the first dominant peak in $p(a^*)$ when a^* is increasing from zero (i.e., leftmost peak). That is,

$$a = \min \arg\{dp(a^*)/da^* = 0\}, \tag{18}$$

where the minimum is taken over all zeros of the equation. If the histogram is unimodal but

asymmetric, the filter tracks the mode and resembles the order-statistics filter.

In connected speech, where peaks cannot be clearly discriminated or the distribution is multi-modal, Eq. (17) can be employed by choosing a value of γ based on the statistics of gap durations. For instance, if gaps constitute approximately 10% of total duration, then $\gamma = 0.1$ would be a reasonable choice. A judicious choice of γ can result in the filter performing like an edge detector, because it captures the transition from larger to smaller values of the time-evolving sequence a_k^*.

The decision strategies, as depicted in Eq. (17) and Eq. (18), were used to validate the model in simulated and real environments (see Results).

In addition to simulations, the MLE approach was validated with real room data. The experimental methods and data analysis procedures are described in the following sections.

A Sound recordings

To validate the MLE method, sound recordings were made in several rooms, building corridors and an auditorium, with the aim of determining their reverberation times. Sound stimuli that were used included 18-tap maximum length (ML) sequences (period length of $2^{18} - 1$), clicks (100 μ s), hand-claps, word utterances (International Phonetic Assoc., 1999), and connected speech from the Connected Speech Test (CST) corpus (Cox *et al.*, 1987). Recordings were made using a

Sennheiser MK-II omni-directional microphone (frequency response 100-20000 Hz). Microphone cables (Sennheiser KA 100 S-60) were connected to the XLR input of a portable PC-based sound recording device (Sound Devices USBPre 1.5). The recorder transmitted data sampled at 44.1 kHz to a laptop computer (Compaq Presario 1700, running Microsoft Windows XP) via a USB link. The sound stimuli, stored as single-channel pre-sampled (44.1 kHz) WAV files, were played through the headphone output of the laptop, amplified by a power amplifier (ADCOM GFA-535II) and presented through a loudspeaker (Analog and Digital Systems Inc., ADS L200e). Data acquisition and test material playback were controlled by a custom-written script in MATLAB (The MathWorks Inc.) using the Sound PC Toolbox (Torsten Marquardt).

Measurement of T_{60} time using Schroeder's method

To validate the estimation procedure, experimentally recorded data from real listening environments were processed using the ML procedure and compared to results obtained from a widely used method developed by Schroeder (1965). The method proposed by Schroeder determines the decay time-constant from the sound decay curve following the cessation of a broad- or narrow-band noise burst. Briefly, if $r(t)$ is the measured decay curve from a single trial, then the mean squared average of the decay curve $s(t)$ over a large number of trials is related to $r(t)$ by

$$\mathbf{E}[s^2(t)] = \int_t^\infty r^2(x)\,dx, \tag{19}$$

where the sound is assumed to switch off at $t = 0$. Schroeder's method, called the backward

integration method, can be applied to a single broad-band channel or to multiple narrow-band channels. The recorded data were filtered offline in ISO one-third octave bands (21 bands with center frequencies ranging from 100-10000 Hz) using a fourth-order Type II Chebyshev band-pass filter with stopband ripple 20 dB down. The output from each channel was processed by the ML procedure and Schroeder's method using Eq. (19). For the broad-band estimation, the microphone output was processed directly using the two methods.

Due to the limited dynamic range of sounds in real environments, Schroeder's method requires the specification of a decay range. The decay ranges normally used are from -5 dB to -25 dB (20 dB range), and from -5 dB to -35 dB (30 dB range). The decay curves in each range were fitted to a regression line using a nonlinear least squares fitting function (function nonlinsq provided by MATLAB). The fitted function was of the form $A\,a_d^n$, where A is a constant, n is the sample number within the decay window, and a_d is the geometric ratio related to the decay time-constant by $a_d = \exp(-1/\tau_d)$. This is in contrast to the model depicted in Eq. (2) which assumes an exponentially decaying envelope with time-constant τ, whereas Schroeder's decay curve is obtained by squaring the signal. Hence, $\tau_d = \tau/2$. Two estimates of the time-constant were obtained from decay curves fitted to the -5 to -25 dB, and -5 to -35 dB drop-offs. For each fit, the line was extrapolated to obtain T_{60} time (in seconds) using the expression

$$T_{60} = \frac{6}{\log_{10}(e^{-1})\,\log_e(a_d)} = \frac{-6\,\tau_d}{\log_{10}(e^{-1})} = 13.82\,\tau_d. \tag{20}$$

The same procedure was followed for determining the time-constant from the broad-band sig-

nal. It should be noted that the ML procedure does not require the specification of a decay range, but only the specification of the estimation window length; thus, only one estimate per band is obtained.

Verification of MLE procedure with ideal stimuli

Microphone data were processed using the MLE procedure to obtain a running estimate of the decay time-constant. For model verification, estimation was performed on: 1) the segment following the cessation of a maximum-length sequence or a hand-clap, and 2) the entire run of a string of isolated word utterances. These were considered ideal stimuli because they fulfilled the model assumptions of free decay or possessed long gaps between sound segments. The estimates were binned for each run and a histogram was produced. The histogram was examined for peaks, and the time-constant was selected using the order-statistics filter Eq. (18) if there were multiple peaks, or Eq. (17) if the histogram was unimodal. The estimate $\hat{a}$ so obtained was used to calculate T_{60} (in seconds) using the formula

$$T_{60} = \frac{3}{\log_{10}(e^{-1})\,\log_e(\hat{a})} = \frac{-3\,\tau}{\log_{10}(e^{-1})} = 6.91\,\tau. \tag{21}$$

In theory, the T_{60} expressions given by Eq. (20) and Eq. (21) are identical due to the relationship between τ and τ_d. However, the calculated values may differ, and this can be ascribed to either model inadequacies or discrepancies in measurement and analysis.

Verification of MLE procedure for speech

The performance of the MLE was also verified using connected speech played back in a circular building foyer (6 m diameter). Test material were connected sentences from the CST corpus. Estimates from non-overlapping 1 s intervals were binned to yield a histogram, and the first dominant peak from the left of the histogram was selected to determine the room time-constant. The procedure for calculating T_{60} time followed Eq. (21).

The estimation procedure was applied to a variety of data sets, including simulated data and real room responses. To illustrate the methods and identify the strengths and deficiencies of the estimation procedure, we first consider simulated data sets. Subsequently we will provide results for real data that validate the room time-constant estimates, and compare these to results from Schroeder's method.

Broad-band white noise bursts in simulated rooms

A 100 ms burst of broad-band white noise (8 kHz bandwidth) was radiated into a simulated room having a decay time-constant $\tau = 100$ ms (Fig. 3). Room output shown in the bottom trace of

Fig. 3 shows the characteristic rise and decay of sound following onset and offset of noise burst (horizontal bar). The graph shows the running estimate of decay time-constant obtained in a 200 ms time window by advancing every sample. Time frames up to about $t = 0.3$ s are not regions of free decay, and so the estimator tended to produce values of $a > 1$. When this was observed in the root-bracketing step of the estimate, the root-solving procedure was aborted. Thus all estimates of a were bounded above by 1. It can be seen that when the window crosses into the region of free decay, the estimator output stabilizes at the true value (horizontal dashed line). A histogram of the time-constant estimates (right axis) was input to the order statistics filter Eq. (17) with $\gamma = 0.5$. The reported time-constant from the filter was $\tau = 101$ ms.

For comparison, the procedure was repeated with the simulated noise burst input (i.e., before it was convolved with the room impulse response) to mimic anechoic conditions. The histogram of a^* demonstrated a strong peak at $a = 1$ ($\tau = \infty$) (not shown). This showed that in the absence of reverberation, as in an anechoic environment or open space, histograms showing strong peaks at $a = 1$ are to be expected.

Effect of window length on estimation

A parameter that is critical for estimation performance is the window length N specified in Eq. (8). Small window lengths are expected to increase the variance of the estimate, as also indicated by the Cramer-Rao lower bound (Eq. 15). To test the effect of window length a burst of white

noise (100 ms duration) was convolved with a simulated room impulse response (τ = 100 ms), and the estimator tracked the decay curve using four different window lengths. The results are shown in Fig. 4. As window length increased from 0.5τ to 4τ, the MLE procedure gave improved estimates. Further, for all four window lengths, there was no bias in the estimates of the peak position. We concluded that increasing window length reduced the variability in the estimates, and did not introduce significant bias.

Although it is desirable to have long window lengths, in practice this is limited by the duration and occurence of gaps between sound segments in the room output. Ideally the filter length should be of the order of τ or longer, but if the gaps are short, then increasing the filter length beyond the mean gap will produce undesirable end effects where the next sound segment creeps into the window. Thus, the window length should not be less than one-half or one-third of τ, but the upper limit is dictated by the mean duration of gaps.

Speech sounds in simulated room

The examples considered above illustrated the performance of the algorithm when the input was broad-band white noise. To be applicable in realistic conditions, the algorithm must perform in a variety of conditions and with different signal types. Speech represents an example where the algorithm is expected to perform poorly, because it is nonstationary and nongaussian. Further, the offset transients in speech sounds (including plosives) have a decay time that can vary from 5-40

ms (unpublished observations). Thus, estimation of decay times with speech presents a particular challenge to the algorithm. We took a sequence of 15 distinct and isolated American-English words recorded in an anechoic environment at a sampling rate of 20 kHz (International Phonetic Assoc., 1999). These included eleven consonant-vowel-consonant words (/p,b,g/V/d/, e.g., "bed"), and four consonant-vowel words (/b/V/, e.g., "bay") separated by a mean interval of 200 ms. These were convolved with a simulated room impulse response having time-constant $\tau = 100$ ms. The task of the estimator was to track the decays for the entire duration of the sequence (approximately 11.4 s). The control condition was the clean input (i.e., anechoic). The results are shown in Fig. 5. Four different filter lengths were used as in Fig. 4. For the control condition (left column) no reliable estimates were produced for the smallest three windows (top three panels) because the histogram peaked at values of τ approaching ∞. For the simulated room response (right column), the peak shifted towards the true value of τ, with the best estimates being obtained for the largest window size of 4τ (right column, bottom row). In all the histograms the peak was located at about 115 ms (arrow). This estimate deviated from the real time-constant of 100 ms due to the lack of sharp transients in the clean speech. A gradual sound offset tends to prolong the reverberated sound even further. This can be seen in the "anechoic" control condition where a small peak is noticeable when window size is 4τ (bottom panel, left column). The peak occurs around 60 ms, and corresponds to the gradual offsets of speech sounds. Thus, this introduces a bias in the estimates under reverberant conditions.

The results of the preceding sections demonstrate the importance of selecting a suitable estimation window length. The choice of window length determines the variability of the estimates, and

is critical to obtaining a histogram with a clearly resolved peak at the true value of the room time-constant. However, the effect of variability on the order-statistics filter is difficult to determine as the filtering operation is nonlinear. Further, bimodal or multimodal histograms may be obtained if there is fluctuating background noise or if the sound segments have an intrinsic offset decay rate (as shown in Fig. 5).

The effect on estimation of offset decay in speech

The preceding section introduced the problem of estimating the room decay time-constant when the input signal exhibited varying offset decays. Here we examine in greater detail the performance of the estimator with input comprising a single word (/b/V/, "bough"). The word was recorded under anechoic conditions and presented to the estimator without modification so that the effect of the vowel offset could be determined. The results are shown in Fig. 6. The terminating vowel has a gradually decaying offset (top panel). Estimation of the offset decay was performed from $t = 0.45$ s (vertical dashed line)using two procedures. First, the envelope was extracted from the analytic signal via a Hilbert transform, windowed, and filtered to eliminate frequency components above 100 Hz. The envelope is shown in the middle panel (heavy outline). The envelope was then squared and transformed to a decibel scale, and the decay time-constant was estimated in windows of duration 0.4 s (horizontal bar), using a least squares fit to a straight line. Successive estimates were obtained by sliding the window forward in steps of one sample. Note that the time

at which an estimate is reported for any given window is the end point of the window. The estimate for the window indicated by the horizontal bar, for instance, is plotted at time $t = 0.85$ s. A curve of the estimated time-constants was thus obtained (dotted curve, bottom panel). The MLE procedure was applied to the same segments and produced an independent estimate of the decay time-constant (solid line, bottom panel). While the estimates differ somewhat, they are in qualitative agreement. Both procedures indicate that the terminating vowel had a time-dependent decay rate, and the greatest rate was between 50 and 70 ms.

The results confirm the presence of the peak in Fig. 5 (left column, bottom panel), although the histogram shown in Fig. 5 was obtained for a sequence of 15 words. The analysis shown in Fig. 6 also indicates the reason for estimation bias under reverberant conditions using speech samples. The offset decays present in clean speech segments will be convolved with the room response, and the estimated time-constants will consequently be greater than the room time-constant. Taken together, the results from Figs. 4-6 suggest that the factors responsible for estimation performance are: the presence of adequate numbers of gaps, sharp offset transients, and estimation window length.

Validation of method

The above results demonstrate that estimation of decay time in room response is possible for a variety of sounds including impulses, noise bursts and speech. Although we have shown that a reasonable agreement exists with a nonlinear least squares fit to the data (Fig. 6), a more careful

evaluation is necessary to determine the conditions under which the MLE procedure is likely to provide accurate estimates. Here we establish that the estimated decay times are comparable to decay times obtained from Schroeder's method (Schroeder, 1965). Furthermore, any data collected must be under sufficiently realistic conditions where there is background noise and where the testing sound is not subject to experimental control. A comparison of MLE performance with the standard method in real environments will therefore establish the utility of the method.

We compared the estimates using Schroeder's method (Schroeder, 1965) in both single-channel (i.e., the broad-band signal), and multi-channel frameworks (i.e., narrow-band signals occupying ISO one-third octave bands). Schroeder's method requires a fitting procedure to estimate the time-constant in a pre-selected decay range (either 20 or 30 dB below a reference level of -5 dB, see Section III). The MLE procedure does not require the specification of such a range.

To determine whether the two methods provide the same RT value, estimations were performed on a simulated room decay curve with RT = 0.5 s (Fig. 7. Broad-band and one-third octave band estimates were obtained using the MLE method (circle) and Schroeder's method (20 dB: lozenge, 30 dB: square). Figure. 7A shows the mean value of RT as a function of center-frequency of the one-third octave bands (open symbols) and the broad-band estimate (filled symbols near y-axis). range) averaged over 100 trials. The broad-band estimates were 0.504 s (MLE), and 0.5 s (Schroeder's method) for both 20 and 30 dB decay ranges. While the MLE estimate was significantly different from Schroeder's method ($p < 0.0001$, Wilcoxon rank sum test), the discrepancy was not greater than 1%. The one-third band MLE estimates in most cases were somewhat

higher than the Schroeder estimates by about 0.5% (mean RT over all bands were, MLE: 0.505, Schroeder's method: 0.502 s for 20 dB and 0.501 s for 30 dB). However, the estimates were not significantly different ($p > 0.0001$), except for one estimate obtained from the 30 dB edcay curve in the band centered at 8 kHz. The most noticeable difference between the two methods was in the variability of the estimates as measured by the standard deviation over the trials (Fig. 7B). The MLE method demonstrated lower SD across trials than Schroeder's method, by a factor of 2 (for the 20 dB curve) and 3 (for the 30 dB curve). Further, MLE estimates were similar across one-third octave bands at frequencies above 200 Hz, (Fig. 7A), whereas estimates from Schroeder's method exhibited greater variability. The results establish that the MLE method and Schroeder's method are in good agreement when tested on model data. While the MLE method may over-estimate the RT when using broad-band signals (although this is no more than 1%), the narrow-band estimates are comparable to those obtained from Schroeder's method, are consistent over a wide range of frequencies, and subject to less variability.

We first report on the comparison between the methods using a hand-clap in a small office (8x3 m). Subsequently we will summarize results obtained in rooms of different sizes. Figures 8A, B depict a hand-clap event and its spectrogram, respectively. The data in panel A is the same as shown in Fig. 1A, except that Fig. 8A also includes the direct sound. The RMS noise level in the room was 50 dBA SPL, and the peak sound pressure level resulting from the hand-clap was 85 dBA SPL. The decay curve obtained using Schroeder's method is shown in Fig. 8C, nor-

malized so that the peak SPL was 0 dB. This is the broad-band curve obtained by integrating the recorded microphone signal. A straight-line fit to the 20 dB drop-off point (circle) from a reference level of -5 dB (lozenge) yielded $\tau = 56$ ms ($T_{60} = 0.39$ s). The discrepancy between this value and that presented in Fig. 1 ($\tau = 59$ ms) was due to the inclusion of the direct sound in Fig. 8. The windows over which the 20 dB drop-off was computed were not identical for the two cases. The data were run through the MLE procedure and a histogram of estimates was obtained, and the decay time-constant was calculated from the peak of the histogram using Eq. (17). This gave an estimate $\tau = 53$ ms ($T_{60} = 0.37$ s), which is in good agreement with the estimate obtained using Schroeder's method. Note that the estimates reported in this work are based on a single trial. The normal practice is to average over large numbers of trials. However, our goal is to develop an online estimation procedure, and so we felt that it would be more realistic to use a single trial.

To test a range of room RTs, ISO one-third octave band analysis (exceeding 1 kHz center frequency) was performed in three environments. These were (1) the moderately reverberant room described above (Fig. 8), (2) a highly reverberant circular foyer, and 3) a highly reverberant enclosed cafeteria. In all cases, the signal was a hand-clap generated at a distance of 2 m from the recording microphone (peak value 90 dB SPL). Output from the band-pass filters were analyzed using the MLE procedure, and the τ value for each band was obtained from the histogram by selecting the dominant peak. For Schroeder's method, a 20 dB decay range was used. Figure 9 shows the T_{60} estimates from Schroeder's method (abscissa) versus the ML estimates (ordinate) for each

ISO one-third band (open symbols), and the average over these bands (closed symbols).

Figure 9 shows that the variability of estimates for highly reverberant environments increases with increasing mean RT for both methods. However, the two methods are in good agreeement especially in the high-frequency bands (the single outlier falling below the diagonal in Fig. 9 is the lowest center frequency used in the analysis, namely 1 kHz). The agreement between the methods is best when the T_{60} values are averaged over all bands (filled symbols), as is usually reported in the literature.

A more extensive test to determine the variability in estimates across different environments, and between bands, was performed in 12 environments, including small office rooms, an auditorium, large conference rooms, corridors, and building foyers. The data were analyzed as in Fig. 9 and are shown in Fig. 10A. In comparison with Schroeder's method, the MLE procedure consistently overestimated T_{60} in low to moderately reverberant environments ($T_{60} < 0.3$ s) whereas it underestimated the reverberation time for more reverberant environments ($T_{60} > 1.3$ s). There was a good agreement between the two methods for intermediate ranges. The average T_{60} over all bands (filled squares) were, however, in good agreement. Broad-band estimates were made using the same procedures but without band-pass filtering of recorded signals. These are shown in Fig. 10B. The trend in the estimates was similar to that observed with narrow-band signals, except

for one outlier. The outlier along with three other data points were obtained in a large auditorium. The latter three were obtained with a source-to-microphone distance of 1.5 m, whereas for the outlier the distance was 4 m. The sound levels were not adjusted to compensate for the distance, and hence the test corresponding to the outlier was at a lower SPL, resulting in reduced dynamic range (from peak SPL to noise floor). For these tests, the Schroeder estimates of T_{60} (in seconds) were 2.18 (outlier), and 0.39, 0.39, and 0.33, respectively. The ML estimates, on the other hand, were 0.69 (outlier), 0.77, 0.80, and 0.67, respectively. Schroeder's method appeared to be inaccurate due to the reduced dynamic range. On the other hand, the ML estimates, while larger than the Schroeder estimates, were consistent and relatively robust to the reduction in dynamic range.

These results raise the issue of estimation in narrow bands. It appears, although it is by no means conclusive, that the upper one-third octave bands (over 1 kHz) may provide more accurate estimates than the lower bands. Frequency decomposition is a standard part of most audio signal processing algorithms, it may be useful to track estimates in the higher frequency bands, or in select bands where the energy is greatest. Tracking high-energy bands is likely to provide more temporal range in tracking decays before encountering the noise floor, and thus sharpen the peak in the histogram of estimates. Alternatively, averaging over all high-frequency bands can provide estimates that are in closer agreement with T_{60} times obtained from Schroeder's method.

The findings suggest that there is good correlation between the estimates obtained with the MLE procedure and those obtained with Schroeder's method. While it is not possible to determine which method provides greater accuracy, we suggest that the values are correlated and in general

agreement.

Estimation of RT from connected speech in real listening environments

The results presented in the preceding sections indicate that the ML estimator output is in good agreement with actual or simulated room RTs. In particular, the estimator can be applied to isolated word utterances, even though the naturally decaying offsets of terminating phonemes may lead to an over-estimation of RT (see Fig. 6). Here, we test the performance of the procedure explicitly in a challenging estimation task, namely estimating room RT from connected speech.

A segment of speech (about 50 s in duration) from the Connected Speech Test (CST) corpus was played back in a partially open, circular foyer (one-third octave band analysis shown in Fig. 9, square symbols). The RT for this environment was first estimated with hand-claps using Schroeder's method (1.66 ± 0.07 s) and independently confirmed with the ML procedure (estimated RT from histgoram was 1.62 s). The ML procedure was then applied to the recorded speech data (Fig. 11A). A histogram of room time-constants for the duration of the recorded data was constructed (Fig. 11B). The order-statistics filter was used to select the first dominant peak in the histogram (RT = 1.83 s). This is the best RT estimate based on the aggregate data. It is possible to refine the procedure for arriving at the best estimate by applying the order-statistics filter at much shorter time intervals. Towards this end, a histogram was constructed at intervals of 1 s, and the best RT estimate for this interval was obtained. The resulting best estimates from all one-

second durations (50 in all) were binned to produce the histogram shown in Fig. 11C. It can be seen that the number of estimates peaks at RT = 1.7 s, which agrees with the mean value of 1.66 s from Schroeder's method (using hand-claps), and is well within its standard deviation (0.07 s; the one-sigma interval is indicated by the horizontal bar in Fig. 11B,C).

Given that terminal phonemes have a natural decay rate (see Fig. 6), it is not surprising that the ML procedure produces estimates somewhat larger than the real room RT. Further, the discrepancy between the actual RT and those estimated from connected speech arise from the absence of adequate numbers, and the limited duration, of gaps (see Fig. 11A). Thus, regions of free decay where estimation is accurate are limited. Not withstanding these constraints, the procedure works well, in part due to the decision-making capability built into the order-statistics filter. By selecting the first dominant peak (from the left) in the histogram, the filter in effect rejects spurious estimates, thereby reducing the error in the estimation procedure. The mean value of the histogram or its median, for instance, would result in significantly higher estimates of RT. The performance of the order-statistics filter can be further improved if one were to obtain a statistical characterization of gap duration from a large corpus of connected speech or other sounds. Such a characterization can provide a robust percentile cut-off value (see Eq. 17) which could then be used to select the best RT value for the room (results not shown).

In conclusion, the ML procedure, in combination with order-statistics filtering, provides a robust means for blind estimation of room RT. The procedure has been validated against Schroeder's method, and with real room data such as hand-claps, isolated word utterances, and connected

speech.

The estimation of reveberation time is a widely investigated problem. Traditionally, two approaches have been taken. The RT is computed analytically using formulae that incorporate the geometry and absorptive characteristics of the reflecting surfaces, or a test sound with known properties is radiated into the environment, and the RT is estimated from the received sounds. The former approach is embodied in the Sabine-type formulae (Sabine, 1922; Eyring, 1930; Millington, 1932; Sette, 1933; see Young, 1959; Kuttruff, 1991, for reviews), while the latter is based on Schroeder's decay curve analysis (Schroeder, 1965; Chu, 1978; Xiang, 1995). In both approaches, prior knowledge of the environment or test sound is required. For example, in Sabine-type formulae, the volume, surface area and absorption coeffiecients must be known; and in Schroeder's decay curve analysis, the test sound must be uncorrelated noise that is abruptly switched off at a known time, and followed by a sufficiently lengthy pause to track the decay. Thus, the methods are not suited for RT estimation from passively received sounds in unknown environments (i.e., blind estimation).

The MLE procedure gets around these difficulties as it is based on a widely accepted model of the reverberant tail, namely the exponential decay model (see Young, 1959, for a discussion on how the Sabine type formulae are related to a linear decay of the sound pressure level after the source is turned off). Here, it is assumed that the amplitude of successive reflections are damped

exponentially, while the fine structure is a random uncorrelated process. This is a good approximation of reverberation in most diffusive environments, and so the method presented here provides a framework for blind estimation with wide applicability. The success of the approach also derives from the analytically tractable nature of the maximum-likelihood formulation, reducing the problem to the estimation of a single parameter that can be determined computationally. We also showed that for ongoing and onset segments of the sound, the estimates will assume implausible values as the model is not valid in these regions. However, an order-statistics filter downstream to the ML estimation can reject these estimates and extract the room RT with improved confidence. This is based on the intuitive idea that sounds cannot decay faster than the rate prescribed by the room time-constant, and thus selecting the earliest peak improves the confidence of the estimates. To our knowledge, this approach has not been reported in the literature.

The two encouraging results of this study are the results obtained using speech sounds and the validation of the estimates using Schroeder's method. Speech sounds present particular problems to most estimation algorithms because they violate the two most commonly held assumptions, namely stationarity and Gaussian statistics. Further, even abruptly terminating phonemes such as stop consonants demonstrate a decay at the cessation of the sound. Such decays may be in the range of 5-40 ms and can increase the overall decay time estimated in reverberant environments. However, except for the increase in estimated decay time (a variation up to about 15% for sounds terminating in /d/) the tracking and histogram procedure works rather well, indicating that the method is relatively robust to model uncertainties.

Partially blind approaches to RT estimation have previously been described. (1) A neural network can be trained to learn the characteristics of room reverberation (Nannariello and Fricke, 1999; Cox *et al.*, 2001). Here, it is necessary to train the network whenever the environment changes. (2) The signal is explicitly segmented to identify gaps wherein decays can be tracked (Lebart *et al.*, 2001). It should be noted that the order-statistics filter developed in this work performs an implicit segmentation of the signal by rejecting estimates that are implausible. (3) A blind dereverberation procedure can be used to obtain the room impulse response. However, the room impulse response must be minimum phase, a condition that most listening environments fail to satisfy (Neely and Allen, 1979; Miyoshi and Kaneda, 1988).

The ML procedure presented here is just one method for estimating room RT. Other methods are also possible. For instance the envelope of the sound can be extracted in the estimation interval, converted to sound pressure level, and a regression line could be fitted to obtain the T_{60} time. This is a blind version of the RT estimation procedure followed by Lebart et al. (2001). The order-statistics filter can be applied to the histogram of estimates as with the ML procedure. The method is non-parametric and so is not subject to model uncertainties. This approach was used to estimate the decay rate of isolated word utterances (Fig. 6). While a detailed comparison of the methods is beyond the scope of this work, we note that the MLE procedure is a principled way of extracting the decay rate from the sound envelope.

The ML procedure is model-based and is expected to perform reasonably well in diffuse sound fields (i.e., uniform with respect to directional distribution) and where a single time-constant de-

scribes the reverberant tail. For most sound fields this is a resonable approximation (see Kuttruff, 1991, for a discussion on this point). The estimates of T_{60} are in good agreement with Schroeder's method in most of the listening environments tested, including challenging situations where the source or recording microphone was close to a wall, or there was moderate background noise (see Fig. 8). Further, it provided consistent estimates even when the dynamic range of sound decay was reduced. In contrast, Schroeder's method seems to provide inaccurate estimates under these conditions (see Fig. 10 and accompanying text). While the ML procedure produces best results when there are isolated impulsive sounds or abruptly terminating white noise bursts, the results of tests with isolated word utterances and connected speech are in good agreement with the actual T_{60}. Thus, the procedure is expected to work under most listening conditions.

The method proposed here can be expected to perform poorly when there are room resonances and the sound pressure level decays nonlinearly with time. This can be a result of the room geometry, or positioning the recording microphone in a region of the sound field that is nondiffusive (e.g., against a reflecting surface). In addition to model failure, the performance of the estimator may be poor when there are insufficient numbers of gaps, or there is fluctuating background noise. Good performance results when there are about 10% gaps and the peak sound level (at the time of offset) is about 25 dB SPL over the noise floor. Performance may also be compromised when background noise is modulated (such as with background music or babble) as the procedure will attempt to track any modulation present in the environment, and hence produce multi-modal histograms with peaks that may not be easily discriminated.

The blind estimation procedure suggested here can be applied in a number of situations. Because only passive sounds are used, any audio processor that has access to microphone input can estimate the room reverberation time, either in single-channel (broad-band) or multi-channel (narrow-band) mode. Further, while the method presented here is for a single microphone, it can be applied with no modifications to an array of microphones, providing several independent estimates of the RT. One of the most interesting applications is in the selection of signal processing strategies tailored to specific listening environments. These include hearing aids and hands-free telephony. Programmable hearing aids often have the ability to switch between several processing schemes depending on the listening environment (Allegro *et al.*, 2001). For instance, in highly diffusive environments, where the source-to-listener distance exceeds the critical distance, adaptive beamformers are ineffective (Greenberg and Zurek, 2001). In such situations, it would be convenient to switch off the adaptive algorithm and revert to the relatively simple (fixed) delay-and-sum beamformer. Alternatively, in highly-confined listening environments such as automobile interiors, where a reflecting surface is located in close proximity to the ear, it may be convenient to switch-off the proximal ear microphone, and use the input from the microphone located in the better (more distal) ear. Such decisions can be made if there is a passive method for determining reverberation characteristics. Other potential applications could include hands-free telephony, and room acoustics evaluation in sound-level meters. A limitation of the method is its relatively poor performance with narrow-band signals whose center frequencies are below 1 kHz. However, the performance is good for broad-band signals, and narrow-band signals whose center frequncies exceed 1 kHz.

The computational costs of implementing the procedure are largely due to the iterative solution of the maximum-likelihood equation. We have developed fast algorithms for reducing the computational cost so that the procedure can be implemented in real-time (forthcoming publication). Thus, the method can be implemeted in passive listening devices to determine the reverberation characteristics of the environment.

All publications and patent applications cited in this document are herein incorporated by reference as if each individual publication or patent application were specifically and individually indicated to be incorporated by reference, including, but not limited to the following:

Attias, H., Schreiner C.E. (1998), "Blind source separation and deconvolution: The dynamic component analysis algorithm," Neural Comp. 10: 1373-1424.

Bell, A.J., Sejnowski, T.J. (1995), "An information maximization approach to blind source separation and blind deconvolution," Neural Comp. 7: 1129-1159.

Cremer, L., Muller, H. (1978), "Principles and Applications of Room Acoustics", T. Schultz (Transl.,) Vol. I. London: Applied Science.

Allegro, S., Buechler, M., and Launer, S. (2001), "Automatic sound classification inspired by auditory scene analysis," Proc. European Conf. Sig. Proc., EURASIP.

Bolt, R. K., and MacDonald, A. D. (1949), "Theory of speech masking by reverberation," J. Acoust. Soc. Am. **21**, 577-580.

Chu, W. T. (1978), "Comparison of reverberation measurements using Schroeder's impulse method and decay curve averaging method," J. Acoust. Soc. Am. **63**, 1444-1450.

Cox, R. M., Alexander, G. C., and Gilmore, C. (1987), "Development of the connected speech test (CST)," Ear and Hear. **8**, 119-126.

Cox, T. J., Li, F., and Darlington, P. (2001), "Extracting room reverberation time from speech using artificial neural networks," J. Audio Eng. Soc. **49**, 219-230.

Eyring, C. F. (1930), "Reverberation time in "dead" rooms," J. Acoust. Soc. Amer. **1**, 217-241.

Greenberg, J. E., Zurek, P. M. (2001), "Microphone-array hearing aids," in *Microphone Arrays: Signal Processing Techniques and Applications,* edited by M. Brandstein and D. Ward (Springer-Verlag, Berlin), pp. 229-253.

Hartmann WM (1997). "Listening in a room and the precedence effect," in *Binaural and Spatial Hearing in Real and Virtual Environments,* edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum Associates, Inc.), pp. 191-210.

International Phonetic Association (1999). *Handbook of the International Phonetic Association* (Cambridge University Press, Cambridge). The American-English sound files are available online at <http://uk.cambridge.org/linguistics/resources/ipahandbook/american-English.zip>.

Knudsen, V. O. (1929), "The hearing of speech in auditoriums," J. Acoust. Soc. Amer. **1**, 56-82.

Kuttruff, H. (1991), *Room Acoustics* (Elsevier Science Publishers Ltd., Lindin, 3rd ed.).

Kuttruff, H. (1995), "A simple iteration scheme for the computation of decay constants in enclosures with diffusely reflecting boundaries," J. Acoust. Soc. Amer. **98**, 288-293.

Lebart, K., Boucher, J. M., and Denbigh, P. N. (2001), "A new method based on spectral subtraction for speech dereverberation," Acustica **87**, 359-366.

Millington, G. (1932), "A modified formula for reverberation," J. Acoust. Soc. Amer. **4**, 69-82.

Miyoshi, M., and Kaneda, Y. (1988), "Inverse filtering of room impulse response," IEEE Trans. Acoust. Speech and Sig. Proc. **36**, 145-152.

Nabalek, A. K., and Pickett. J. M. (1974), "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," J. Speech Hear. Res. **17**, 724-739.

Nabalek, A. K., and Robinson, P. K. (1982), "Monaural and binaural speech perception in reverberation for listeners of various ages," J. Acoust. Soc. Amer. **71**, 1242-1248.

Nabalek, A. K., Letowski, T. R., and Tucker, F. M. (1989), "Reverberant overlap- and self-masking in consonant identification," J. Acoust. Soc. Amer. **86**, 1259-1265.

Nannariello, J., and Fricke, F. (1999), "The prediction of reverberation time using neural network analysis," Appl. Acoust. **58**, 305-325.

Neely, S. T., and Allen, J. B. (1979), "Invertibility of room impulse response," J. Acoust. Soc. Amer. **66**, 165-169.

Pitas, I., and Venetsanopoulos, A. N. (1992), "Order statistics in digital image processing," Proc. IEEE **80**, 1893-1921.

Poor, V. (1994), *An Introduction to Signal Detection and Estimation* (Springer–Verlag, New York).

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipes in C* (Cambridge Univ. Press, Cambridge).

Sabine, W. C. (1922), *Collected Papers on Acoustics* (Harvard University Press, Cambridge).

Schroeder, M. R. (1965), "New method for measuring reverberation time," J. Acoust. Soc. Amer. **37**, 409-412.

Schroeder, M. R. (1966), "Complementarity of sound buildup and decay," J. Acoust. Soc. Amer. **40**, 549-551.

Sette, W. J. (1933), "A new reverberation time formula," J. Acoust. Soc. Amer. **4**, 193-210.

Tahara, Y., and Miyajima, T. (1998), "A new approach to optimum reverberation time characteristics," Appl. Acoust. **54**, 113-129.

Xiang, N. (1995), "Evaluation of reverberation times using a nonlinear regression approach," J. Acoust. Soc. Amer. **98**, 2112-2121.

Young, R. W. (1959), "Sabine reverberation equation and sound power calculations," J. Acoust. Soc. Amer. **31**, 912-921

Any theory, mechanism of operation, proof, or finding stated herein is meant to further enhance understanding of the present invention and is not intended to make the present invention in any way dependent upon such theory, mechanism of operation, proof, or finding. While the invention has been illustrated and described in detail in the figures and foregoing description, the same is to be considered as illustrative and not restrictive in character, it being understood that only selected embodiments have been shown and described and that all changes, modifications and equivalents that come within the spirit of the invention as defined herein or as follows are desired to be protected.

One embodiment of the present invention includes a unique technique for evaluating reverberation. Other embodiments include unique methods, systems, devices, and apparatus to determine reverberation time of a room.

A further embodiment includes a sensor for detecting sound and a processing subsystem. The processing subsystem receives sound-representative signals from the sensor to determine reverberation time of an unknown acoustic environment by processing these signals with a maximum-likelihood estimator. In one form, processing further includes applying an order-statistics filter.

In yet a further embodiment of the present invention, a system includes a sensor and a processor. The processor is responsive to signals from the sensor to evaluate reverberation by estimating one or more reverberation characteristics in accordance with a maximum likelihood function and applying an order-statistics filter.

In still a further embodiment, a memory device includes instructions executable by a processor to evaluate reverberation based on sound-representative signals from a

sensor. The instructions define a routine to iteratively estimate one or more reverberation characteristics based on a maximum likelihood function and order-statistics filter. In one form, the memory device is removable and is of a disk, cartridge, or tape type.

Another embodiment of the present invention is directed to a method that comprises: detecting sound with a sensor to generate a corresponding sensor signal; generating data with the sensor signal in accordance with a maximum likelihood estimator; and filtering data with an order-statistics filter to provide an estimate of reverberation time. In one form, processing is performed within a single wideband channel spanning a frequency range of interest. In another form, processing is performed with respect to each of a number of narrowband channels; where the narrowband channels each correspond to a different acoustic signal frequency range. For this form, the estimate can be determined by combining estimation results for each of the channels. A further embodiment is a system, device, or apparatus operable to implement this method in one or more of its forms.

Yet another embodiment of the present invention includes a method, system, device, and/or apparatus to determine reverberation time of an acoustic environment. This determination includes iteratively determining at least two values corresponding to a maximum likelihood function to evaluate one or more reverberation characteristics of an acoustic environment. In one form, one of the values corresponds to a time-constant parameter and/or another of the values corresponds to a diffusive power parameter of the reverberation. The evaluation can be completely or partially blind. In one partially blind approach, a neural network is included and trained to provide the reverberation characteristics of a selected room. outside region, or other acoustic environment based on

the maximum likelihood evaluation. In addition, various filtering and windowing operations can be performed to further evaluate reverberation, such as applying an order-statistic filter to estimates form the maximum likelihood evaluation, processing sound data over one or more selected frequency bands, and/or adaptively changing processing window lengths.

In still another embodiment, a reverberation evaluation system, method, apparatus, or device of the present invention is provided with a hearing aid or other hearing assistance device, a cochlear implant, a hands-free telephony arrangement, a speech recognition arrangement, a telepresence/teleconference configuration, and/or sound level evaluation equipment.

Abstract

The reverberation time (RT) is an important parameter for characterizing the quality of an auditory space. Sounds in reverberant environments are subject to coloration. This affects speech intelligibility and sound localization. Many state-of-the-art audio signal processing algorithms, for example in hearing-aids and telephony, are expected to have the ability to characteristize the listening environment, and turn on an appropriate processing strategy accordingly. Thus, a method for characterization of room RT based on passively received microphone signals represents an important enabling technology. Current RT estimators, such as Schroeder's method, depend on a controlled sound source, and thus cannot produce an online, blind RT estimate. Here, a method for estimating RT without prior knowledge of sound sources, or room geometry is presented. The diffusive tail of reverberation was modeled as an exponentially damped Gaussian white noise process. The time constant of the decay, which provided a measure of the RT, was estimated using a maximum-likelihood procedure. The estimates were obtained continuously, and an order-statistics filter was used to extract the most likely RT from the accumulated estimates. The procedure was illustrated for connected speech. Results obtained for simulated and real room data are in good agreement with the real RT values.

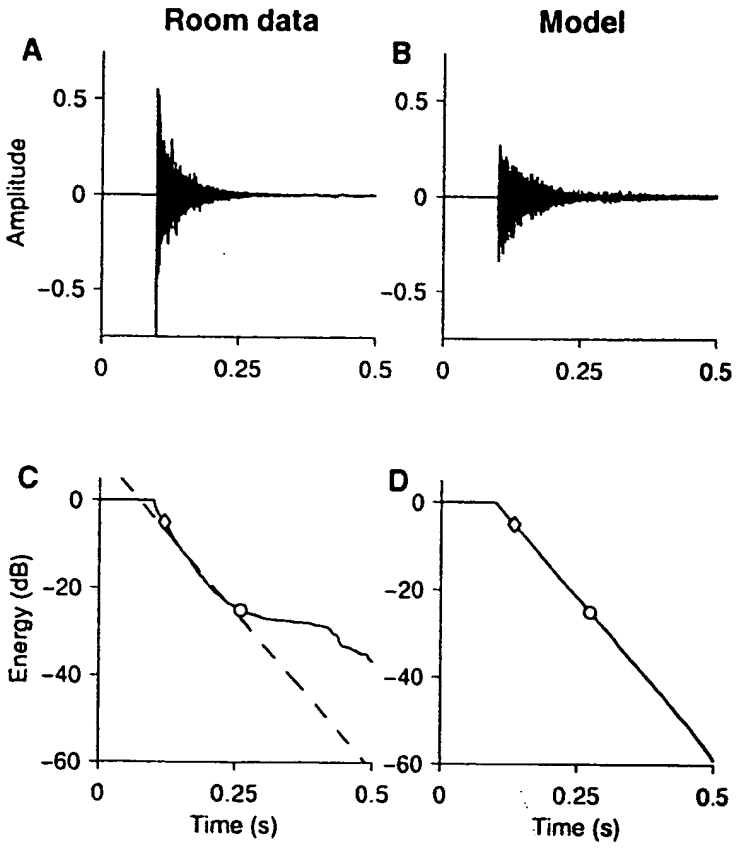

Room data
Model
A
B
C
D
Amplitude
0.5
0
-0.5
0
0.25
0.5
Energy (dB)
-20
-40
-60
Time (s)

Figure 1:

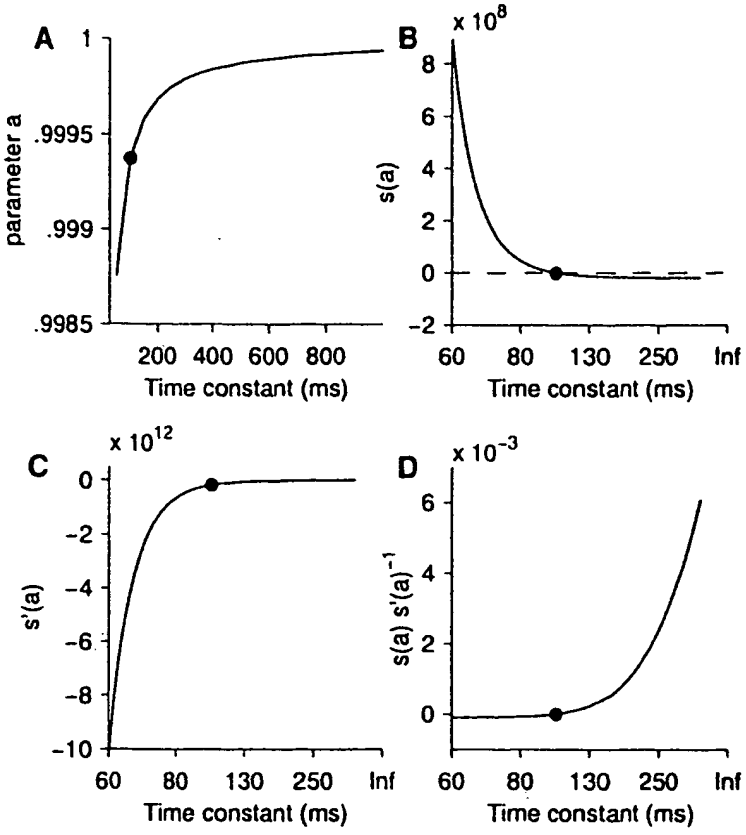

A
parameter a
1
.9995
.999
.9985
200 400 600 800
Time constant (ms)
B
x 10^8
s(a)
8
6
4
2
0
-2
60 80 130 250 Inf
Time constant (ms)
C
x 10^12
s'(a)
0
-2
-4
-6
-8
-10
60 80 130 250 Inf
Time constant (ms)
D
x 10^-3
s(a) s'(a)^-1
6
4
2
0
60 80 130 250 Inf
Time constant (ms)

Figure 2:

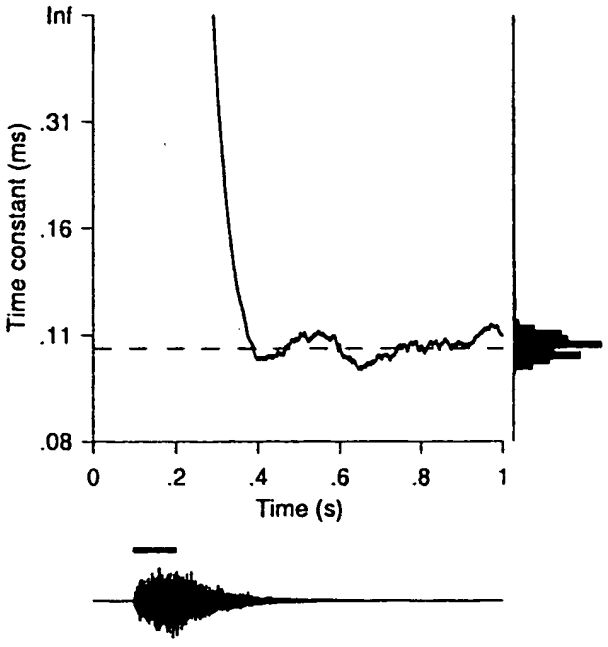


Figure 3:

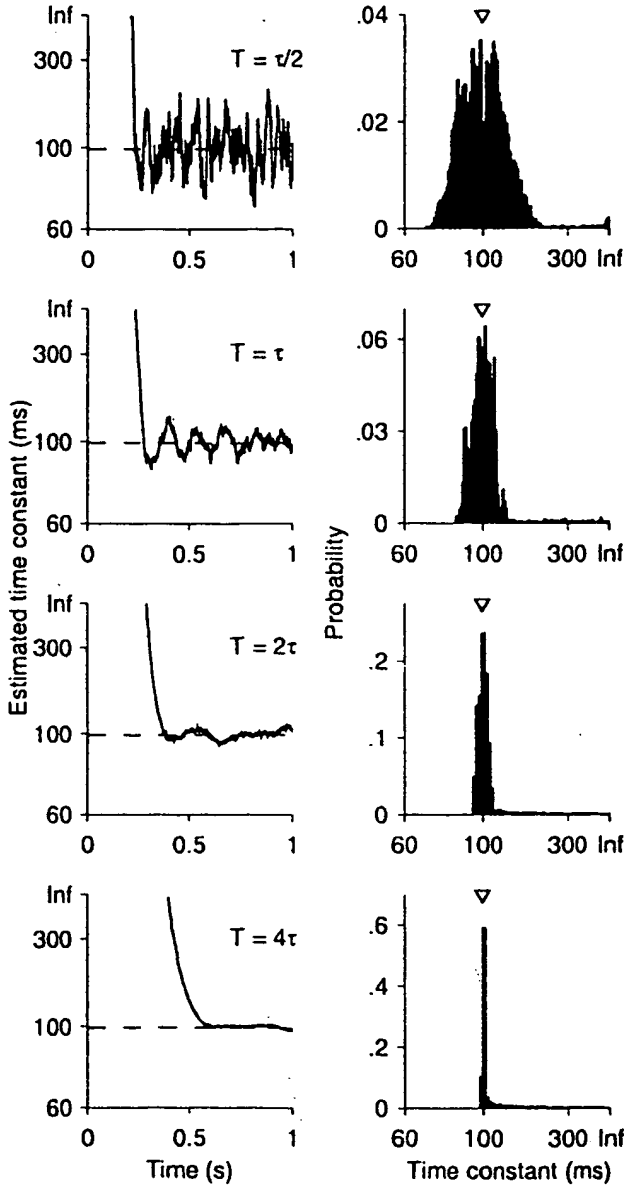

Inf
300
100
60
T = τ/2
T = τ
T = 2τ
T = 4τ
0
0.5
1
Time (s)
Estimated time constant (ms)
.04
.02
.06
.03
.2
.1
.6
.4
Probability
60
100
300
Time constant (ms)

Figure 4:

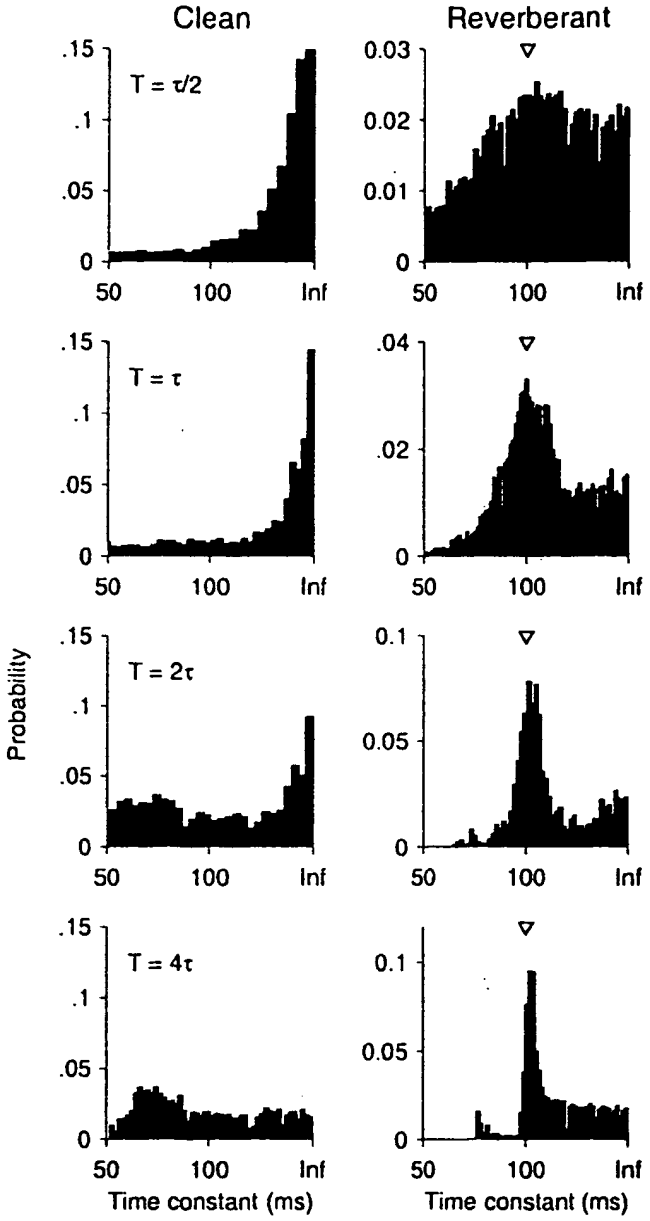

Clean
Reverberant
T = τ/2
T = τ
T = 2τ
T = 4τ
Probability
Time constant (ms)
Time constant (ms)
50
100
Inf

Figure 5:

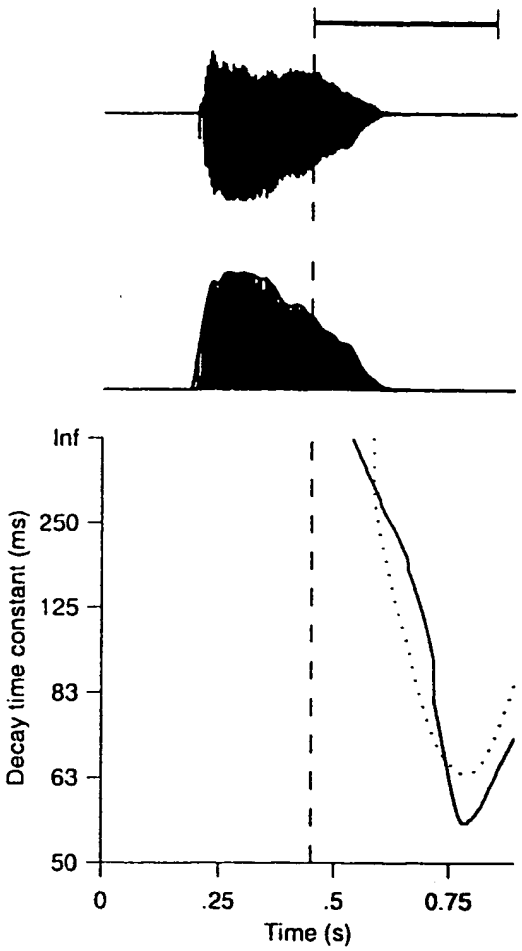


Figure 6:

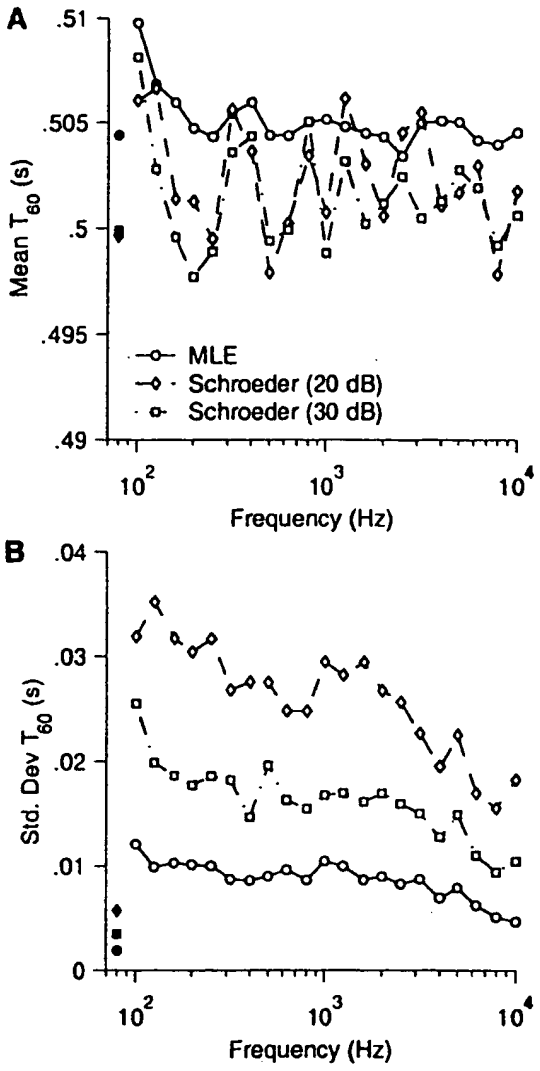

A
.51
.505
.5
.495
.49
Mean T_{60} (s)
MLE
Schroeder (20 dB)
Schroeder (30 dB)
10^2
10^3
10^4
Frequency (Hz)
B
.04
.03
.02
.01
0
Std. Dev T_{60} (s)
10^2
10^3
10^4
Frequency (Hz)

Figure 7:

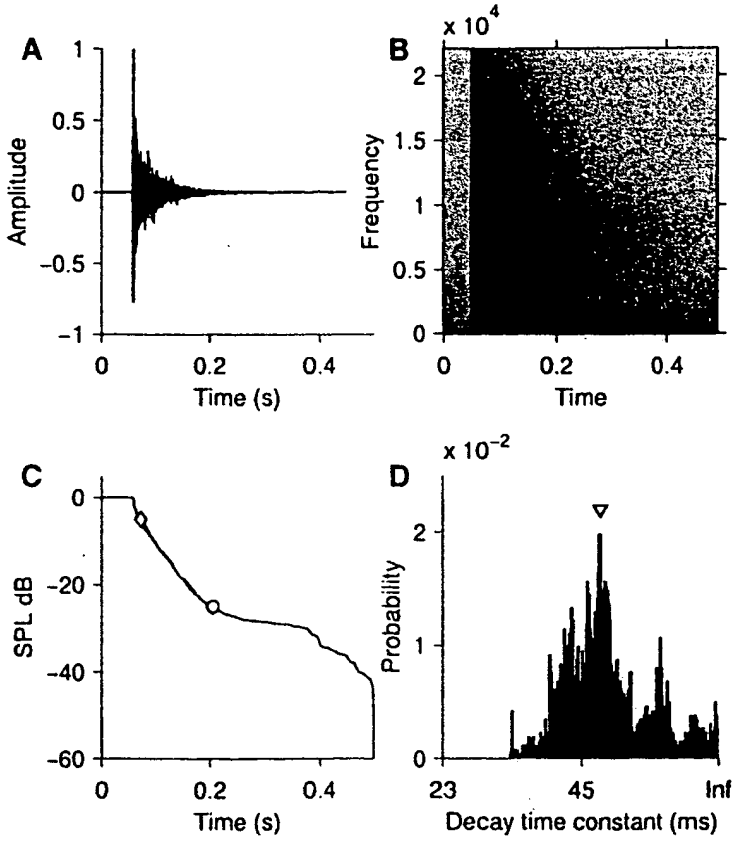

A
Amplitude
1
0.5
0
-0.5
-1
0
0.2
0.4
Time (s)
B
x 10^4
Frequency
2
1.5
1
0.5
0
0
0.2
0.4
Time
C
SPL dB
0
-20
-40
-60
0
0.2
0.4
Time (s)
D
x 10^-2
Probability
2
1
0
23
45
Inf
Decay time constant (ms)

Figure 8:

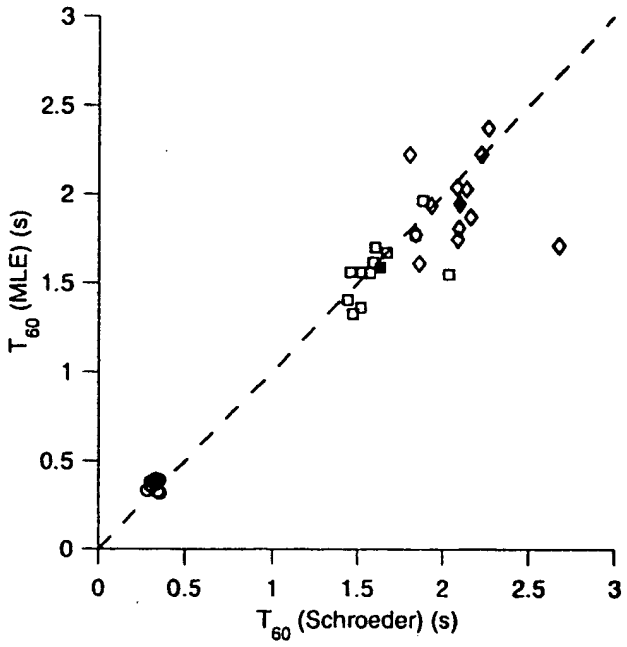


Figure 9:

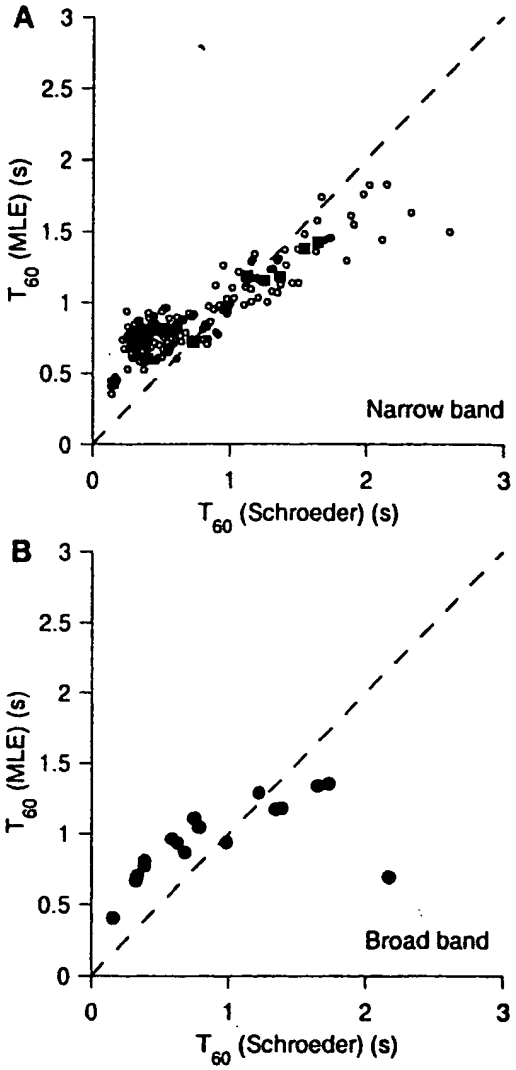

A
B
T_{60} (MLE) (s)
T_{60} (Schroeder) (s)
Narrow band
Broad band
0
0.5
1
1.5
2
2.5
3

Figure 10:

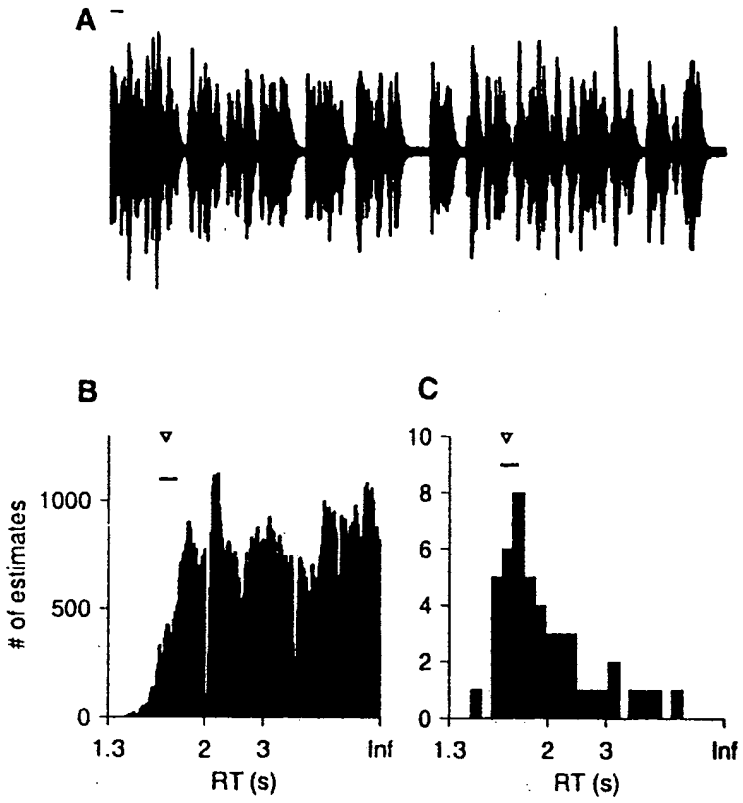


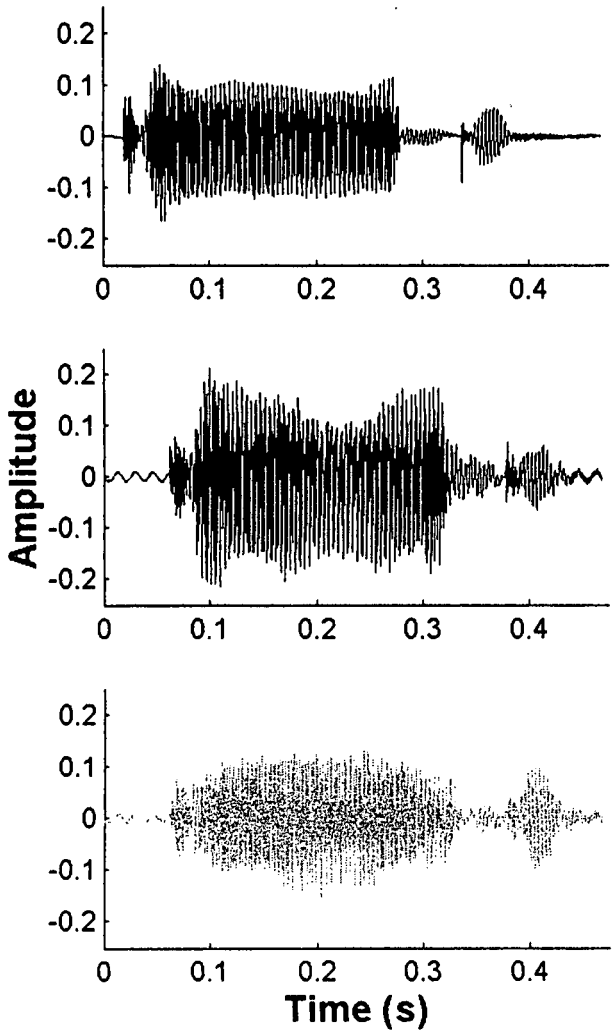
Figure 11:

Reverberation

- Delayed repeated copies of a sound are added (i.e., convolution and not addition)
- Distorts envelope and fine structure of sound
- Blurs localization cues
- Raises overall sound intensity and can improve intelligibility
- Can provide a sense of spaciousness

Fig. 12

The Degradation of Speech



Dry speech

- Fine structure / envelope preserved
- Stop closures and release noticeable

Right Mic

- Fine structure blurred
- Envelope preserved
- Stop closures show decay

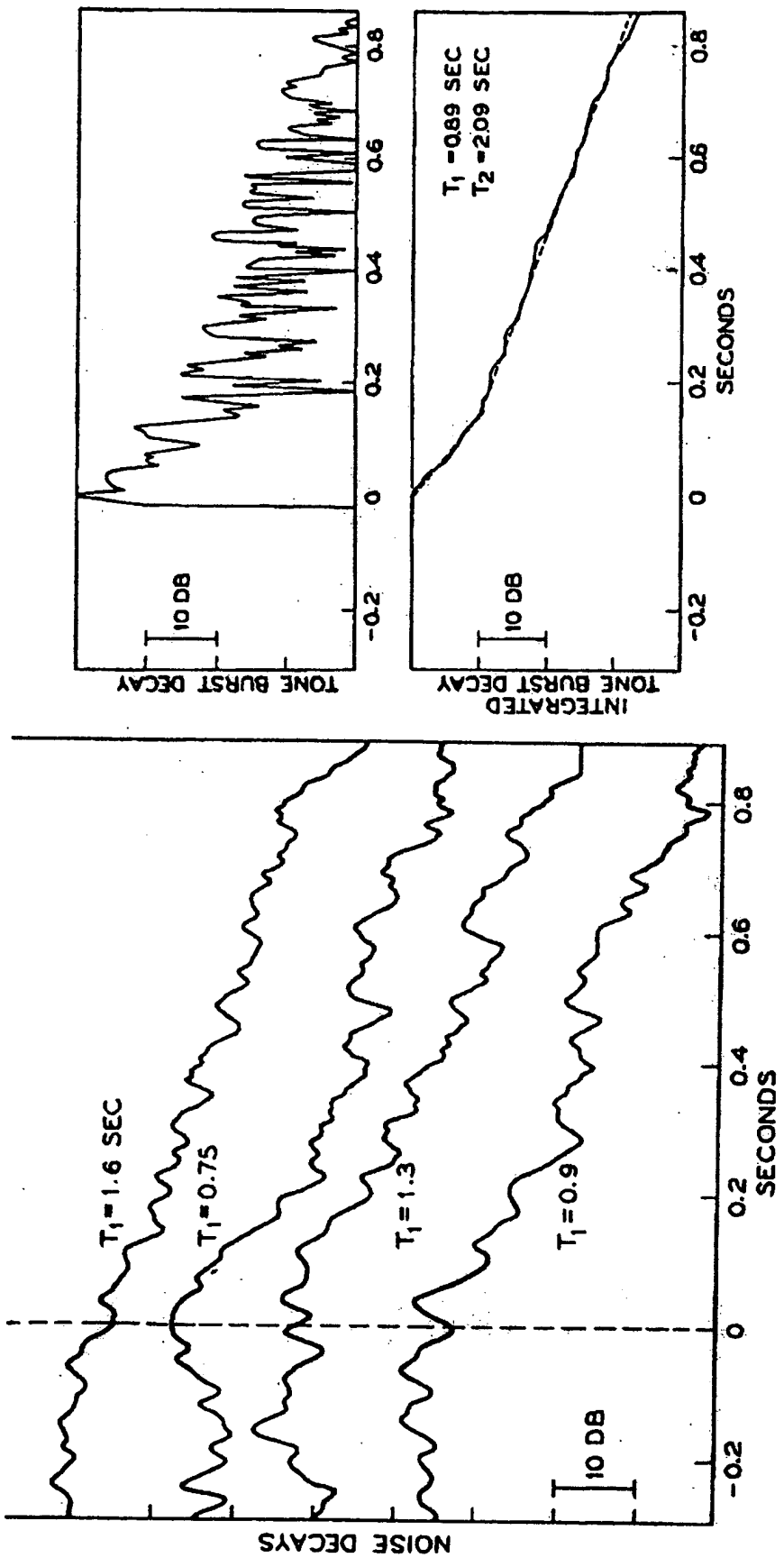
Left Mic

- Fine structure/envelope blurred
- Envelope shows build-up/decay
- Stop closures show decay

Fig. 13

REVERBERATION TIME EVALUATION
Inventors: Rama Ratnam et al.
App. No. NEW Atty Docket No. 22010-200
Drawing sheet 13 of 24

Decay curves in a room



From: Schroeder (JASA 1966)

Fig. 14

An Estimator for Tracking Free Decays

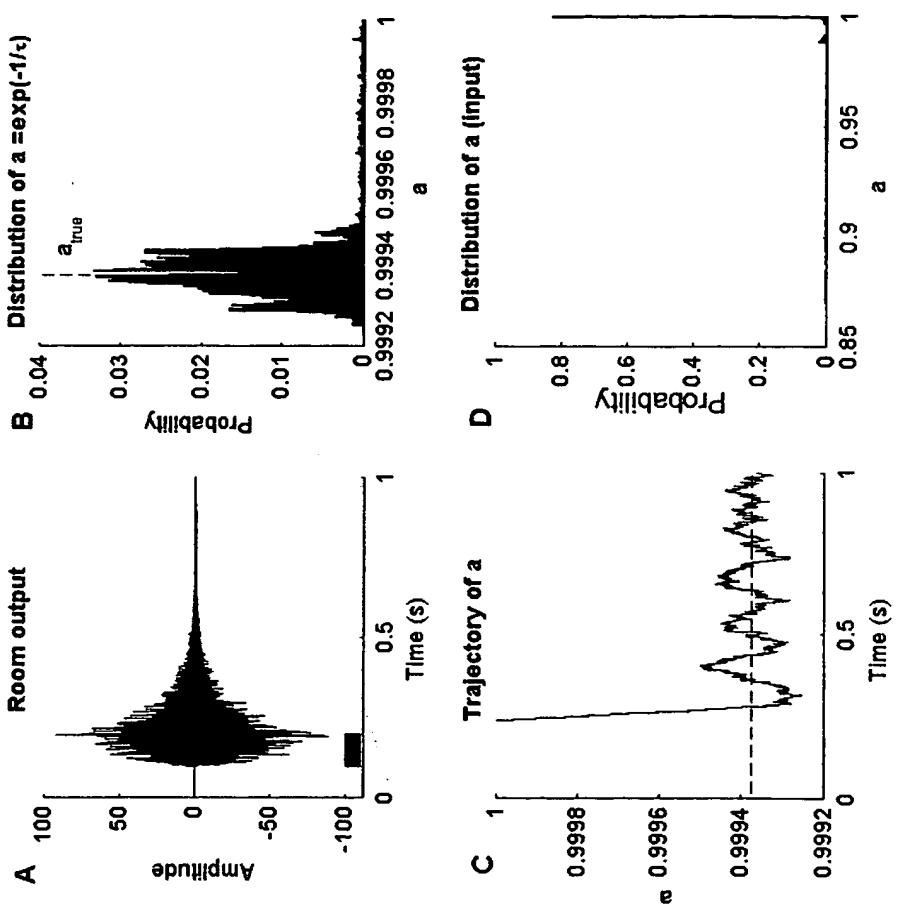


Fig.1 7

Decision making strategy

- There is variability in a even during free decays
- The estimates are implausible during ongoing or onset transients
- Hypothesis is that the majority of the estimates at realistic values of a (< 1) will be at the true a
- Employ an order-statistics filter to find the dominant peak in the histogram of a

Fig. 18

Effect of Filter Length on Estimate

- Small windows increase variability
- Large windows may span gaps

→ Compromise

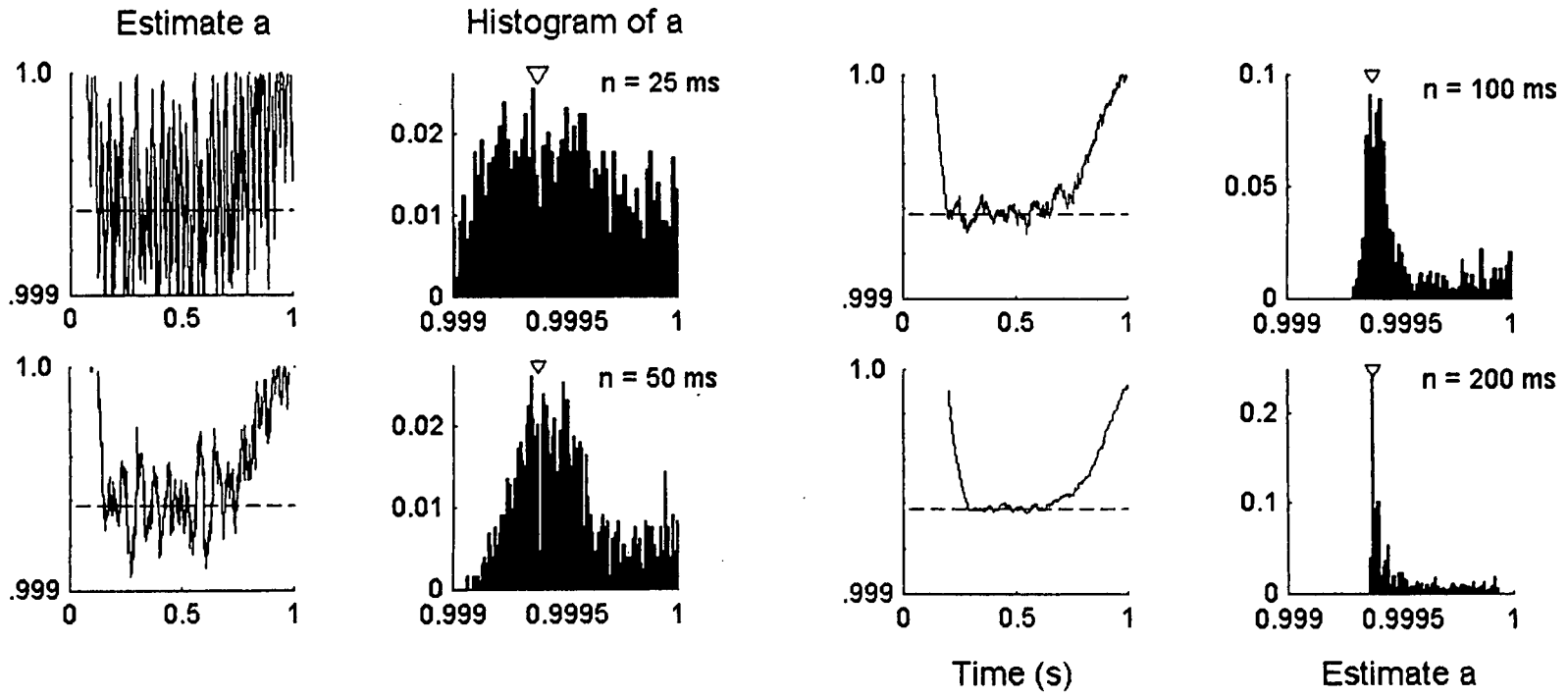
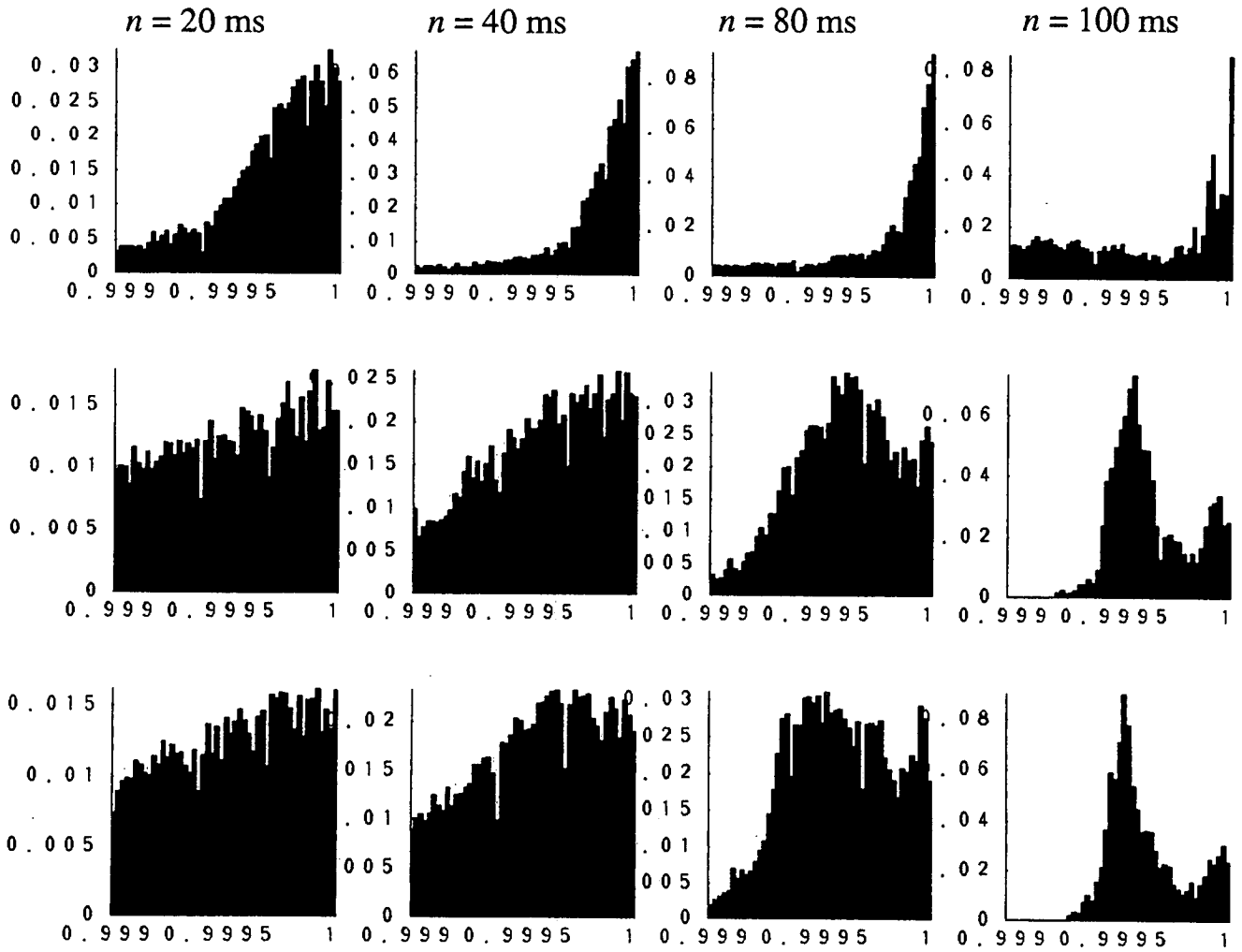


Fig. 19

Tracking free decays in speech (/b/v/d/)



Clean

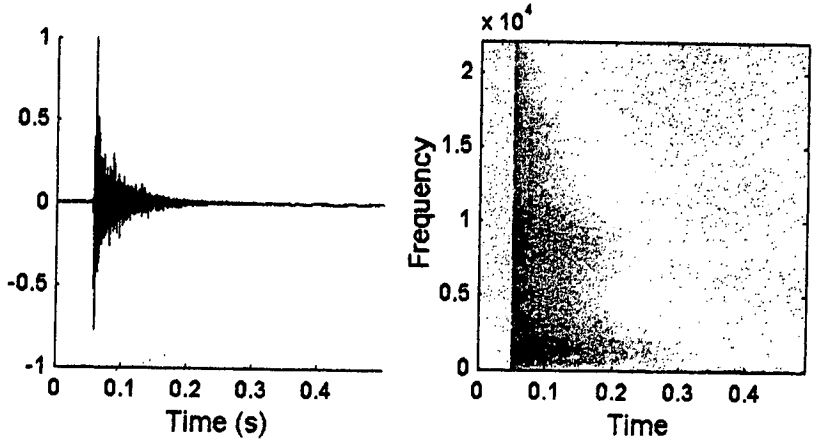
T_{60} =0.5s

T_{60} =0.5s
Randomized
fine structure

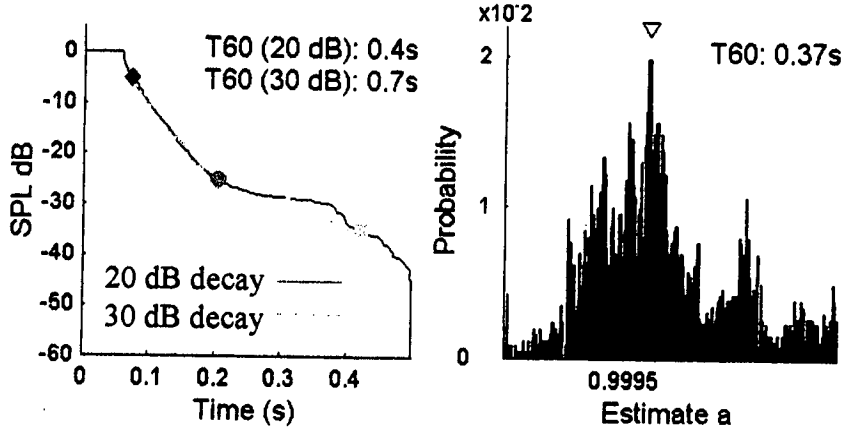
Fig. 20

Validation of method

Hand clap



Schroeder method and regression fit



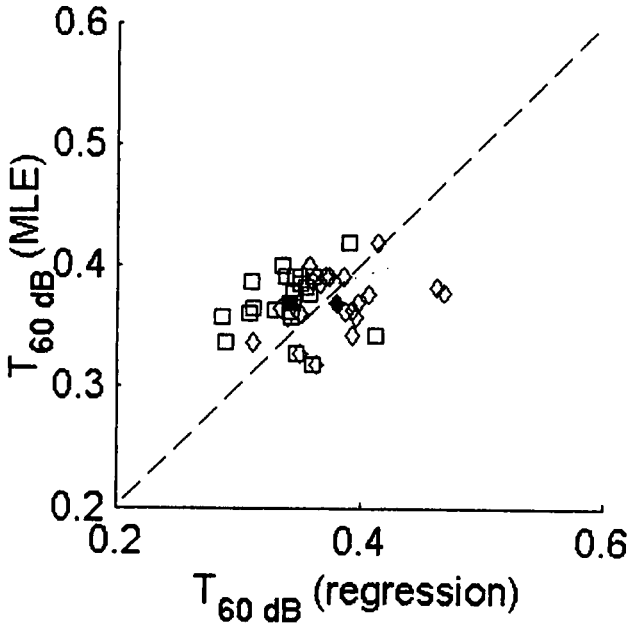
Blind MLE method

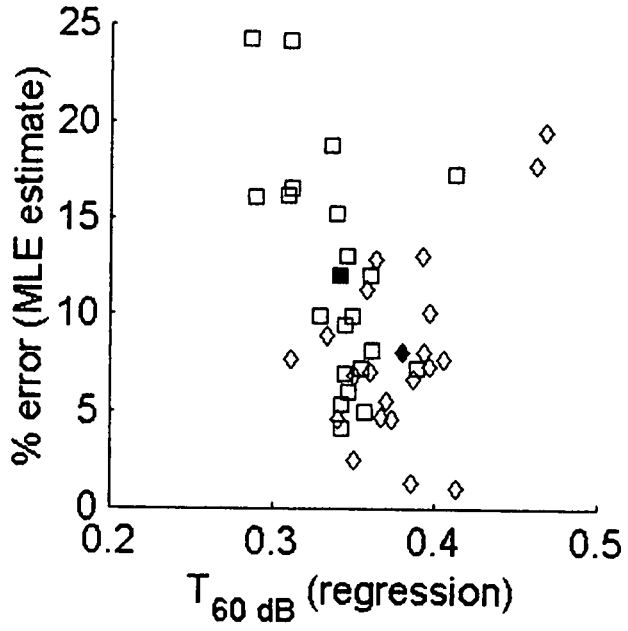
Fig. 21

REVERBERATION TIME EVALUATION
Inventors: Rama Ratnam et al.
App. No. NEW Atty Docket No. 22010-200
Drawing sheet 21 of 24

1/3-Octave Band Estimates of T60 dB:

Comparison of MLE estimator with regression fit to data





Squares: Estimate from 20 dB decay
Lozenge: Estimate from 30 dB decay
Filled: Mean over all bands

Fig. 22

RT Estimation

- The model is validated and performs well given the assumptions
- A sub-band approach is recommended since estimation accuracy improves in high frequency bands (> 1kHz)
- Filter length may be adjusted either adaptively or several lengths used in parallel

Fig. 23

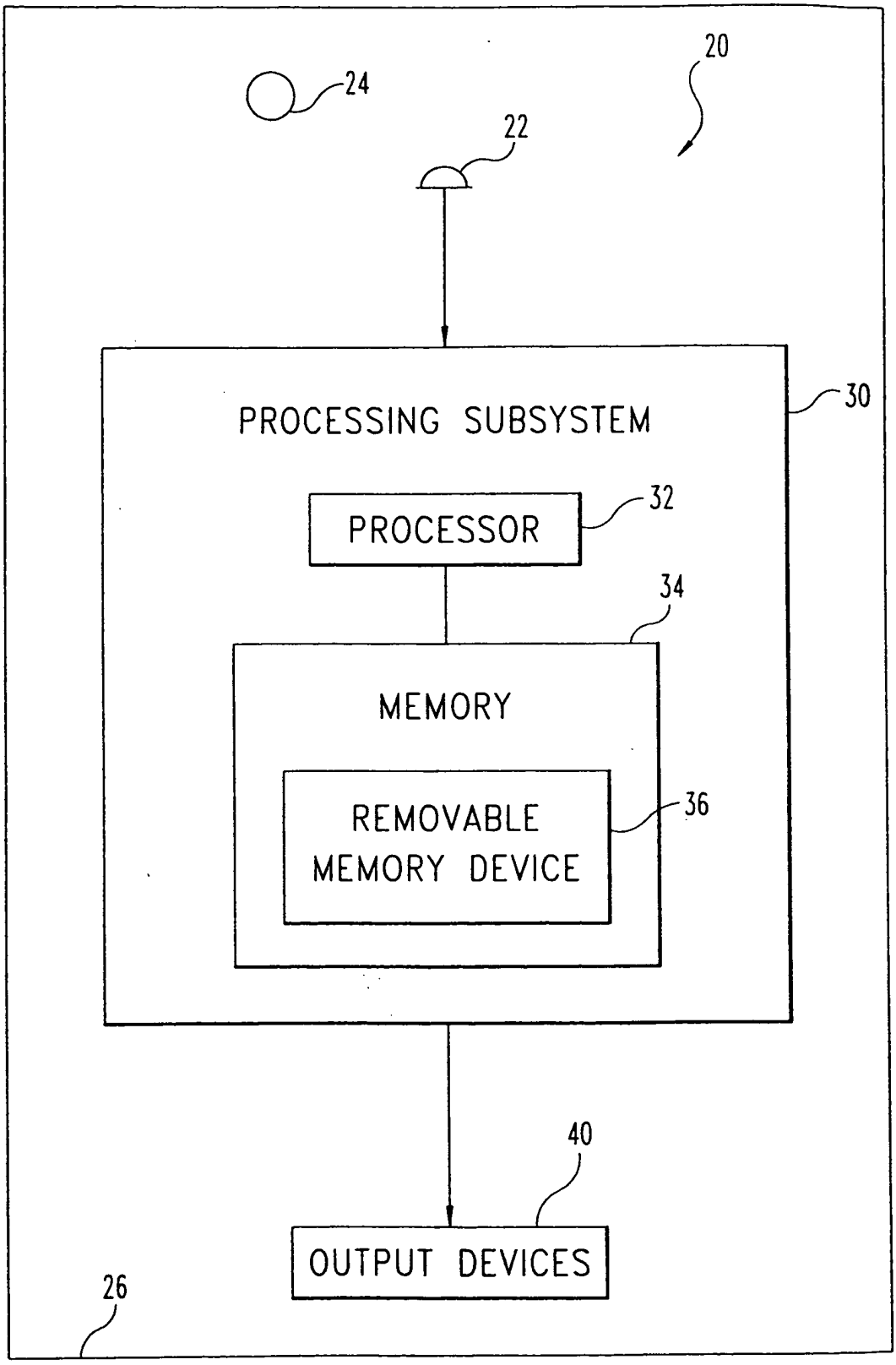

20
24
22
PROCESSING SUBSYSTEM
30
PROCESSOR
32
34
MEMORY
REMOVABLE
MEMORY DEVICE
36
40
OUTPUT DEVICES
26

Fig. 24